

Università degli Studi di Napoli
Federico II

The Density Valued Data Analysis in a Temporal
Framework:
The Data Model Approach

Carlo Drago

Ph.D Dissertation
Statistics

XXIV Cycle



The Density Valued Data Analysis in a Temporal Framework

The Data Model Approach

Napoli, 30 November 2011

"Non se ne andrà più" gli dico.

"Non se ne andrà più?"

"Nulla se ne andrà più."

... "Perché?" egli dice "Che accade?"

"... Nessuna cosa ora è sola."

"Sarebbe ogni cosa anche tutto il resto?"

"Precisamente. E dov'è una cosa, è anche tutto il resto ..."

... Viene l'infanzia lo stesso; viene la terra intesa come fu con fiori bianchi ch'erano di capperi e sembravano farfalle; vengono come sono alla radio, le città del mondo, Manila e Adelaide, Capetown, S. Francisco, di Cina, di Russia, non mai vedute, e Trieste un pò veduta, e così Madrid, Oviedo, e di più che vedute, principio e infanzia di ognuna, Ninive, Samarcanda, Babilonia.

Che altro?

Certo il papà con gli occhi azzurri.

E la madre. La nonna. ...

... Vengono i cavalli ch'erano da ferrare, idem gli uomini loro, i viandanti, i vecchi barboni, i carrettieri. Le lunghe strade con la polvere, anch'esse, e su di esse il sonno, il fieno, fossi di cicale: tutto quello che è stato, e vuole con ognuno che si perde essere ancora.

E il cielo che fu dell'aquilone?

Il cielo che fu dell'aquilone.

da "Uomini e No" di Elio Vittorini

Acknowledgements

My greatest appreciation and acknowledgement, with many thanks, go to my tutors Professor Carlo Lauro and Professor Germana Scepi, who co-authored some of the works presented here, for their guidance, continuous encouragement, suggestions and help. They are the Alpha and the Omega of the research period I have spent on the Ph.D. Professor Carlo Lauro always ensured that I was pursuing the best possible paths, and energetically guided my thought process with many ideas and suggestions, which I have tried to transform into this work. Professor Germana Scepi, at the same time, constantly oriented me to the best work possible both with patience and with the everyday support needed by a researcher. They did an exceptional work on me and although being fully engaged in their Departmental duties, they always found the time for discussion and the development of my ideas. It was an honour and a privilege to know and to work with them.

Also a thanks to Ian Chadwick for the formidable proofreading work he did in the last days of the thesis and to Antonio Balzanella for his tremendous LaTeX skills. I am very grateful for their support.

I have in this sense to acknowledge the entire PhD Committee (Collegio dei Docenti) who gave various relevant and important suggestions for my work: Giuseppe Giordano, Maria Gabriella Grassia, Roberta Siciliano, Simona Balbi, Rosanna Verde, Marina Marino, Marco Gherghi, Francesco Palumbo, Cristina Davino, Beniamino Di

Martino, Antonio Irpino, Massimo Aria.

I would also like to thank all the Professors of the PhD in Statistics for their encouragement and support, and the Professors of the Department of Mathematics and Statistics in general. In particular, Vincenzo Aversa and Sergio Scippacercola.

An important part of this work was developed under the supervision of Professor Carlos Maté when I was in Madrid at Universidad Pontificia Comillas, Institute for Research in Technology.

I have to acknowledge Professor Efraim Centeno, the Institute Director, and also the Institute for the logistic support given for my research in the period I worked in Madrid.

I have received comments, useful suggestions or had stimulating discussions with Edwin Diday, Francisco De Carvalho, Javier Arroyo, Antonio Munoz, Sara Lumbreras, Carolina Ascanio Garcia, Eugenio Sanchez, Elvira Romano, Federica Gioia, Simona Signoriello, Domenico Vistocco, Alessandra Rosato, Edwin Diday, Oldemar Rodriguez, Paula Brito, Monique Noirhomme, Paulo Teles, Lynne Billard, Domenico De Stefano, Antonio D'Ambrosio, Americo Todisco, Alfonso Iodice D'Enza, Raffaele Miele, Davide Carbonai, Valerio Tutore, Clelia Cascella, Marialaura Pesce, Daniela Nappo, Nikolas Pet-sas, Antonio Forte, Angelo Leogrande, Matteo Ruggeri, Boris Suruç, Elvan Celyan, Francesca Perino and Michela Verardo.

I would like to acknowledge, for some useful discussions and general points of view, Marco Riani, Jaromir Antoch, Neyko Neykov, Marcello Chiodi, Bettina Gruen and the participants at various conferences and workshops, and also the anonymous referees who have read previous versions of the submitted works.

I discussed the issues presented here with various experts in the financial and economic sector: Francesco Manni, Paolo Ronzoni, Andrea Iovene, Gaetano Vecchione, Antonio Scarpati, Alfonso Ponticelli and Antonio Semeraro. Thank you also to Corrado Meglio for his support and encouragement during different periods of the research.

Cristina Tortora, has been a great colleague and friend, with whom I have shared the three years of the PhD. We share both thoughts and academic duties, not only our office! Marcella Marchitelli was an important colleague and friend too in the early stages of my PhD, and also when we have co-authored works. The development of a scientific idea is typically nonlinear so thanks to both Cristina and Marcella for sharing with me some parts of our common path.

Agnieszka Stawinoga, Lidia Rivoli, Nicole Triunfo, Maria Spano, Stefania Spina, Mena Mauriello and Maddalena Giugliano shared with me the doctoral duties and offered useful ideas. Enrico Cafaro was very attentive to my needs related to software, both in academic duties and the thesis development. Thanks to you all.

I have to acknowledge various Professors who have taught me over time: Franco Peracchi, Martino Lo Cascio, Ruggero Paladini, Nicola Rossi and Bernardo Maggi.

I have to acknowledge various students I have tutored jointly with Professor Germana Scepi, and Professor Carlo Lauro during the Ph.D. I wish to thank in particular two of them, who won the prestigious Bloomberg Competition Trade Ideas on quantitative trading: Luna Damiani and Danilo Vigliotta; and then Ciro Novizio, Roberta Migliaccio, Alessia Malizia, Rosario Capriglione, Giuseppe di Meglio, Valentina Costagliola, Chiara Matano, Fabrizio Bottari, Daniela Visone, Ilaria Ariola Fabiana D'Antona, Adriano di Guglielmo, Ottavio Telese, Maria Carannante, Antonio Iaquinto, Rosario Capriglione, Pasquale Buo, Anna Mele, Roberta Leopardi, Enrico Infante, Sonia Esilda Cona and Clementina Maresca.

I had, as well, many friends who helped me: Paolo Santella, Andrea Polo, Giulia Paone, Emiliano Miluzzo, Roberto Ricciuti, Francesco Millo, Flavia Weisghizzi, Livia Amidani Aliberti, Enrico Baffi, Enrico Gagliardi, Mariagrazia Albano; and also in a different way Luca Tarrantelli, Antonia Baratta, Fabio Briguglio and Alessia Latino Quirto. Dimitris Diafas, Enrico Picozzi, Leonardo Dell'Annunziata and fam-

ily, Filippo Casazza and family, Federico Cipolla and family, Antonio Montola, Vincenzo Grillo, Alessandro Montalto and Flavio D'Andria.

Peter Gleason and Nancy Bickmore have helped me to revise parts of the entire manuscript. My parents and my family have supported me continuously during the three years of the thesis. I have to say a big thanks to them for everything.

Any errors are clearly my own.

My final thought is for Domenico. He was not a mere spectator of the work. He participated actively in the entire development. I wish you could have been here to read the finished work.

This work is in memory of my friend Domenico Irace.

...Credo che l'uomo sia maturo per altro, per nuovi, altri doveri. E questo che si sente, io credo, la mancanza di altri doveri, altre cose da compiere... Cose da fare per la nostra coscienza in un senso nuovo da "Conversazione in Sicilia" di Elio Vittorini.

Contents

Introduction	1
I Data: The State of The Art	7
1 The Analysis of Massive Data Sets	9
1.1 Complex Data Sets and Massive Data	13
1.1.1 Characteristics of Complex Data Sets and Mas- sive Data	17
1.1.2 Statistical Methods, Strategies and Algorithms for Massive Data Sets	20
1.2 Analysing data using Aggregate Representations	21
1.2.1 Scalar Data and their Aggregate Representation	22
1.2.2 Sources for Aggregate Representations and Sym- bolic Data	29
1.2.3 Complex Data and Tables of Aggregate Repre- sentations	30
1.3 Aggregate Representations from Time Series	36
1.4 A study simulation on Big Data and Information Loss .	38
1.5 Applications on Real Data	39
1.5.1 The Symbolic Factorial Conjoint Analysis for the Evaluation of the Public Goods	39

1.5.2	Analysing the Financial Risk on the Italian Market using Interval Data	41
2	Complex Data in a Temporal Framework	45
2.1	Homogeneous and Inhomogeneous Time Series	47
2.1.1	Equispaced Homogeneous Data	48
2.1.2	Inhomogeneous High Frequency Data	52
2.1.3	Irregularly Spaced Data as Point Processes	54
2.1.4	Inhomogeneous to Homogeneous Time Series Conversions	57
2.2	Ultra High Frequency Data Characteristics	59
2.2.1	Overwhelming number of observations	60
2.2.2	Gaps and erroneous observations in data	61
2.2.3	Price discreteness	63
2.2.4	Seasonality and Diurnal patterns	65
2.2.5	Long dependence over time	66
2.2.6	Distributional characteristics and Extreme Risks	67
2.2.7	Scaling Laws	67
2.2.8	Volume, Order Books and Market Microstructure	68
2.2.9	Volatility Clustering	69
2.3	Financial Data Stylized Facts	70
2.3.1	Random Walk Models and Martingale Hypothesis	71
2.3.2	Distributional Properties of Returns: Fat Tails	73
2.3.3	Heterogeneity and Structural Changes	73
2.3.4	Non-Linearity	74
2.3.5	Scaling	75
2.3.6	Dependence and Long Memory	75
2.3.7	Volatility Clustering	76
2.3.8	Chaos	77
2.3.9	Cross Correlations Between Assets	77

3	Foundations of Intervals Data Representations	79
3.1	Internal Representation Data and Algebra: intervals . .	81
3.1.1	Probabilistic Arithmetic	81
3.1.2	Interval Data and Algebra	82
3.1.3	Statistical methods for Interval Representations	87
3.1.4	Stochastic Processes and Time Series of Interval-Valued Representations	89
4	Foundations of Boxplots and Histograms Data Representations	91
4.1	Internal Representation Data and Algebra: Boxplots, Histograms and Models	92
4.1.1	Quantile Data and Algebra	92
4.1.2	Histogram Data and Algebra	97
4.2	Statistical Methods Involving Boxplots and Histograms valued data	104
4.2.1	Histogram Stochastic Processes and Histogram Time Series (HTS)	105
4.3	Internal Representations Models	105
4.4	The Data Choice	107
4.4.1	The Optimal Data Choice	107
4.4.2	Conversions between Data	109
5	Foundations of Density Valued Data: Representations	111
5.1	Kernel Density Estimators	112
5.2	Properties of the Kernel Density Estimators	114
5.3	The Bandwidth choice	116
5.4	Density Algebra using Functional Data Analysis	118
5.5	Density Algebra using Histogram Algebra	119
5.6	Density Trace and Data Heterogeneity	119
5.7	Conversions between Density Data and other types of data	121

5.8	Simulation Study: effects of the kernel and the band- width choice	122
5.9	Application on Real Data: Analysing Risk Profiles on Financial Data	123
5.9.1	Analysis of the Dow Jones Index	124
5.9.2	Analysis of the financial crisis in the US 2008-2011	125

II New Developments and New Methods 139

6	Visualization and Exploratory Analysis of Beanplot Data	141
6.1	The Data Aggregation problem	144
6.1.1	High Frequency Data and Intra-Period Variability	147
6.1.2	Representations, Aggregation and Information Loss	149
6.2	From Scalar Data to Beanplot Data	155
6.3	Beanplot Data	157
6.4	Beanplot Time Series (BTS)	161
6.4.1	Beanplot Time Series (BTS): Kernel and the Bandwidth Choice	168
6.4.2	Trends, Cycles and Seasonalities	172
6.5	Exploratory Data Analysis of Beanplot Time Series (BTS)	178
6.6	Rolling Beanplot Analysis	181
6.7	Beanplot Time Series (BTS) and Data Visualization: a Simulation Study	183
6.7.1	Some Empirical Rules of Interpretation	191
6.8	Visualization: comparing the Beanplot time series (BTS) to other approaches	193
6.9	Applications on Real Data	195
6.9.1	Analysing High Frequency Data: the Zivot dataset	195
6.9.2	Application on the US Real Estate Market in 1890-2010	196

6.9.3	Comparing Instability and Long Run Dynamics of the Financial Markets	197
6.10	Visualizing Beanplot Time Series (BTS): Usefulness in Financial Applications	198
7	Beanplots Modelling	203
7.1	Beanplot Coefficients Estimation	206
7.1.1	Beanplots Model Data: the modelling process	208
7.2	Coefficients Estimation: The Mixture Models Approach	213
7.2.1	Choosing the optimal interval temporal	218
7.3	Beanplot Representations by their Descriptor Points	219
7.3.1	Descriptor point interpretation: Some Experi- ments on Simulated and real datasets	225
7.4	Data Tables considering Density representations	228
7.5	From Internal to External Modelling	229
7.5.1	Detecting Internal Models as Outliers	230
7.6	Internal Models Simulation	230
8	Beanplots Time Series Forecasting	233
8.1	Density Forecasting and Density Data	234
8.2	From Internal Modelling to Forecasting	236
8.3	External Modelling (I): TSFA from model coefficient approach	237
8.3.1	Detecting Structural Changes	240
8.3.2	Examples on real data: Forecasting World Mar- ket Indices	240
8.4	External Modelling (II): Attribute Time Series Approach from Coordinates	245
8.4.1	Analysis of the Attribute Time Series Approaches	246
8.4.2	Attribute Time Series Forecasting Models	247
8.4.3	Identification and External Modelling Strategy	248

8.4.4	Examples on real data: Forecasting the Beanplot Time Series (BTS) related to the Dow Jones Market	251
8.5	The K-Nearest Neighbour method	253
8.6	The Forecasts Combination Approach	254
8.6.1	Combination Schemes	255
8.6.2	Optimal weight determination	257
8.6.3	Weight determination by regression	258
8.6.4	Identification of the components to model	258
8.6.5	Identification and implementation of the Hybrid modelling strategy	259
8.6.6	Using Neural Networks and Genetic Algorithm in the modelling process	260
8.7	The Search Algorithm	261
8.8	Crossvalidating Forecasting Models	261
8.9	Extremes and Risk Forecasting	262
8.10	Beanplot Forecasting: Usefulness in Financial Applications	263
9	Beanplots Time Series Clustering	267
9.1	Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach	269
9.1.1	An application on real data: Clustering stocks in the US Market	275
9.2	Internal Modelling and Clustering: the Attribute Time Series Approach	279
9.3	Classical Approaches in Clustering Beanplot Features	280
9.3.1	Application: classifying the synchronous dynamics of the world indices beanplot time series (BTS)	282
9.4	Model Based Clustering and Modern Framework	287
9.5	Feature Model Based Clustering for Beanplot Time Series (BTS)	288

9.5.1	The choice of the temporal windows	291
9.5.2	Application: classifying the synchronous dynamics of the european indices beanplot time series (BTS)	293
9.6	Clustering Beanplots Data Temporally with Contiguity Constraints	298
9.7	Clustering using the Wesserstein Distance	300
9.8	Comparative Approaches: Clustering beanplots from Attribute Time Series	302
9.9	Building Beanplot Prototypes (BPP) using Clustering Beanplot Time Series (BTS)	303
9.10	Sensitivity and Robustness of the Clustering Methods .	303
9.10.1	Ensemble Strategies in Clustering Beanplots . .	307
9.11	Clustering: Usefulness in Financial Applications	307
10	Beanplots Model Evaluation	311
10.1	Internal Modelling: Accuracy Measures	312
10.2	Mixture Models and Diagnostics	313
10.2.1	Application on real data: Evaluating Internal Models: the case of the Mixtures	314
10.3	Forecasting Evaluation Methods	316
10.3.1	Forecasting evaluation procedure	317
10.3.2	Discrepancy Measures	318
10.3.3	Applications on Real data: Evaluating the Mixture coefficients estimation and Forecasting . . .	319
10.3.4	Applications on Real data: Evaluating Forecasting the Dow Jones Index	319
10.4	Clustering Evaluation Methods	320
10.4.1	Internal Criteria of cluster quality	320
10.4.2	External Criteria of cluster quality	323
10.4.3	Computational Criteria	324
10.5	Forward Search Approaches in Model Evaluation . . .	324

10.6 The Internal and the External Model Respecification	325
10.6.1 Application on real data: Model Diagnostics and Respecification	325
11 Case Studies: Market Monitoring, Asset Allocation, Statistical Arbitrage and Risk Management	333
11.1 Market Monitoring	334
11.2 Asset Allocation	342
11.3 Statistical Arbitrage	348
11.4 Risk Management	356
Conclusions and Extensions for Future Research	363
A Routines in R Language	377
B Symbols and Acronyms used in the Thesis	379
B.0.1 Symbols	379
B.0.2 Acronyms and Abbreviations	381
Bibliography	385

List of Tables

1.1	Data Analysis Typologies	35
5.1	Risk profiles: quantiles computed 2007-2009	133
5.2	International Stockmarket Symbols	135
6.1	Bandwidth for various beanplot time series (BTS) 2005-2011	200
6.2	Bandwidth for various beanplot time series (BTS) 2005-2011	201
6.3	BVSP	201
6.4	DJI	201
6.5	GDAXI	202
6.6	FTSEMIB.MI	202
7.1	Internal Representations and Descriptor Points	214
7.2	Coefficients estimation example	217
7.3	highest density regions (hdr)	222
7.4	falpha	222
10.1	Internal modelling evaluation	313
10.2	Internal modelling diagnostics	329
10.3	External modelling diagnostics	330

10.4 Accuracy of the Forecasting Models on the Attribute Time Series	331
11.1 Forecasting results	361

List of Figures

1.1	Global Information created and available storage 2005-2011. See The Economist 2010 [655] and Batini 2010 [65]	13
1.2	Computation Capacity 1986-2007 see Hilbert and Lopez 2011 [361] and McKinsey 2011 [499]	14
1.3	Daily Price Change of S&P Grouped By Year, data from Yahoo Finance. Source: VisualizingEconomics.com	18
1.4	Classical data in a medical data set [82]	23
1.5	Interval data in a mushrooms data set [82]	24
1.6	Histogram data in a Cholesterol data set Gender \times Age categories: (Billard 2010 [82])	25
1.7	The process of transformation from complex data tables into symbolic data or aggregate representation tables [81]	34
1.8	The process of transformation from a relational data table into a symbolic data table[209]	37
1.9	Judgement analysis and the interval symbolic data. Marchitelli 2009 [486] and Drago et al. 2009 [230]	42
1.10	Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])	43
1.11	Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])	43

1.12	Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])	44
2.1	High Frequency Data: Trades and quotes dataset (Galli 2003 [286])	49
2.2	High Frequency Data: Trade durations (Galli 2003 [286])	55
2.3	High Frequency Data: Quote durations (Galli 2003 [286])	55
2.4	High Frequency Data: Point Processes (Hautsch 2007) [348])	56
2.5	Erroneous observations in High Frequency Data: one spike (Browne 2011 [746])	64
2.6	Erroneous observations in High Frequency Data: bid ask gapping (Browne 2011 [746])	65
3.1	Intervals (Revol 2009 [581])	83
3.2	A real interval and its parameters lb (lower bound), ub (upper bound), mid (midpoint), rad (radius) and wid (width) Kulpa 2004 [435]	84
4.1	Comparing Internal Representations	93
4.2	Comparing Complex Internal Representations: Sampled data from a $N(50,2)$ summarized by symbolic variables — González-Rivera G. Carlos Maté (2007) [315]	94
4.3	Boxplot time series (BoTS)	97
4.4	Candlestick time series (CTS)	98
4.5	Differences between Histogram and Boxplot Data. Source SAS 9.2 Support Documentation	99
4.6	Histogram Data	101
4.7	Back to Back Histograms	101
4.8	Histogram Time Series (HTS)	103
4.9	Clipping Histograms (Risk Visualization)	104

4.10	Different types of Symbolic Data (Signoriello 2008 [630])	108
5.1	Kernel density estimation, histogram and rugplot on simulated data	112
5.2	Kernel density estimation: illustration of the kernels (Francois 2011 [280])	115
5.3	Overlapped Kernel density estimations [793]	117
5.4	Effect of the kernel and the bandwidth choice	122
5.5	Effect of the kernel and the bandwidth choice	123
5.6	Effect of the kernel and the bandwidth choice (2)	123
5.7	Effect of the kernel and the bandwidth choice (3)	124
5.8	Effect of the kernel and the bandwidth choice (4)	124
5.9	Density Estimation and Profile Risk Indicator computed	127
5.10	Density Estimation and Profile Risk Indicator year: 2007	128
5.11	Density Estimation and Profile Risk Indicator computed year: 2008	129
5.12	Density Estimation and Profile Risk Indicator year: 2009	130
5.13	Density Estimation and Profile Risk Indicator computed year: 2010	131
5.14	Density Estimation and Profile Risk Indicator computed year: 2011	132
5.15	Radius of the Interval time series (ITS) DJI 1990-2011	132
5.16	Bandwidth for the US Densities computed over the years	134
5.17	Implied Volatility for the US Market (VIX Index Index of volatility expectations (Bloom 2009 [89])	137
5.18	Beanplot Time series (BTS) DJI 1990-2001	138
6.1	Intra-day price data for Microsoft stock (Zivot 2005 [722])	145
6.2	Financial data types with the typical sizes and frequency (Dacorogna et al. 2001) [163]	147

6.3	Financial data models with the data typical sizes and frequency (Dacorogna et al. 2001) [163]	149
6.4	The evolution from the boxplot (Harrell et al. 2011 [340])	156
6.5	Uncertainty in Representations	156
6.6	Simulated beanplot time series (BTS) and turning point identification	162
6.7	Dow Jones Index Beanplot Time Series (BTS) considered for the period 1996-2010	167
6.8	Enhanced density data with first and last observation (in red and blue respectively)	168
6.9	Enhanced density data with first and last observation: DJI 1990-2011 (in red and blue respectively)	169
6.10	Enhanced density data with first and last observation: FTSEMIB.MI 2003-2011 (in red and blue respectively)	170
6.11	Enhanced Beanplot Trend for the centre: DAX 1990-2011 (in red and blue, green respectively open, close, centre)	175
6.12	Enhanced Beanplots and Business Cycle Analysis: 3-Year US Unemployment Rates 1948-2011 (in red and blue, first and last observation)	176
6.13	Enhanced Beanplot and Business Cycle Analysis: 3-Year US Capacity Utilization: Total Industry (TCU) 1967-2011 (in red and blue, first and last observation)	177
6.14	Simulated Beanplot Time Series (BTS) and Kernel Smoothers	180
6.15	Simulated Beanplot Time Series (BTS) and Smoothing Splines	181
6.16	Comparing different objects: an example on a single simulated time series: Drago and Scepi 2009	187
6.17	Comparing Boxplot (BoTS) and Beanplot Time Series (BTS) (Drago Scepi 2009)[237]	188
6.18	Comparing different interval temporal periods Drago Scepi 2009 [237]	189

6.19	High Frequency Microsoft Data 1-15 May 2001 (see Drago and Scepi 2009)	196
6.20	Rolling Beanplots Real Home Price Index 1890-2011 using different windows	197
6.21	Rolling Boxplots Real Home Price Index 1890-2011 using different windows	198
7.1	US Dow Jones differenced time series 1990-2011	208
7.2	Beanplot Time Series (BTS) using different Temporal Intervals on an ARIMA(1,1,0) with a structural change	211
7.3	The Data Analysis Cycle	214
7.4	Internal and external modelling	215
7.5	Kernel Density Estimation: computing the area between z_1 and z_2	220
7.6	Highest Density Regions: BVSP Market (differenced series) year 2010	221
7.7	Bovespa Beanplot Time Series -BTS Y^C attribute time series of the descriptor points 1993-2011 ($n = 20$)	225
7.8	Bovespa Beanplot Time Series (BTS) X^C attribute time series of the descriptor points 1993-2011 ($n = 20$)	226
7.9	Comparing descriptors amongst all the Representation Time Series (ITS, BoTS, HTS, BTS and models)	227
7.10	Table of the parameters of the data models (Signoriello 2008 [630])	229
8.1	DJI - Dow Jones Index	242
8.2	DJI - Dow Jones Index	242
8.3	GDAXI - German Dax Index	243
8.4	MMX - Major Market Index Mexico	243
8.5	SSEC - China Stock Market Index	244
8.6	N225 - Nikkei Index Japan	244

9.1	Amazon (AMZN)	271
9.2	Apple (AAPL)	272
9.3	Goldman Sachs (GS)	272
9.4	Microsoft (MSFT)	273
9.5	Deutsche Bank (DB)	273
9.6	Morgan Stanley (MS)	274
9.7	Bank of America (BAC)	274
9.8	Citigroup (C)	275
9.9	Dow Jones Market	276
9.10	Dow Jones Market 2007–2008	277
9.11	Dow Jones Market 2008–2009	277
9.12	Dow Jones Market 2009–2010	278
9.13	Dow Jones Market 2010–2011	278
9.14	Beanplot Time Series (BTS) Model Based Clustering (1)	294
9.15	Beanplot Time Series (BTS) Model Based Clustering (2)	295
9.16	Beanplot Time Series (BTS) Model Based Clustering (3)	296
9.17	Beanplot Time Series (BTS) Model Based Clustering (4)	297
9.18	Beanplot Time Series (BTS) Model Based Clustering (5)	298
9.19	Beanplot Time Series (BTS) Model Based Clustering (6)	299
9.20	Building Beanplot Prototypes (BPP) from the Beanplot Time Series (BTS)	306
10.1	Beanplot Dow Jones Data 1996-2010 (see Drago and Scepi) 2010	321
10.2	Attribute Time Series	328
11.1	Scalar dataset	335
11.2	Visualization of the time series Dow Jones 1990-2011 .	336
11.3	Visualization of the time series Bovespa (Brazil) BVSP 1990-2011	336
11.4	Beanplot time series (BTS) for the Dow Jones DJI 2001- 2011	337

11.5 Beanplot Time Series (BTS) Cac 40 (France) FCHI 1990-2011	338
11.6 Beanplot Time Series (BTS) Bovespa (Brazil) BVSP 1990-2011	338
11.7 Histogram Time Series (HTS) Dow Jones DJI 1990-2011	339
11.8 Histogram Time Series (HTS) Cac 40 FCHI (France) 1990-2011	340
11.9 Histogram Time Series (HTS) Bovespa (Brazil) BVSP 1990-2011	340
11.10 Interval Time Series (ITS) Dow Jones DJI 1990-2011 .	341
11.11 Interval Time Series (ITS) Cac 40 (France) FCHI 1990- 2011	341
11.12 Interval Time Series (ITS) Bovespa (Brazil) BVSP 1990- 2011	342
11.13 Boxplot Time Series (BoTS) Dow Jones DJI 2001-2011	343
11.14 Boxplot Time Series (BoTS) Cac 40 (France) FCHI . .	343
11.15 Boxplot Time Series (BoTS) Bovespa (Brazil) BVSP 2001-2011	344
11.16 Clustering original time series (2000-2011)	348
11.17 Clustering using coordinates correlation distance dis- similarity matrix	349
11.18 Clustering using coordinates correlation distance aver- age method	349
11.19 Clustering using coordinates correlation distance single method	350
11.20 Clustering using coordinates correlation distance Mc- Quitty method	350
11.21 Clustering using coordinates correlation distance Com- plete method	351
11.22 Clustering using correlation distance - Centroid method	351
11.23 Factor 2: correlation distance: (average)	352
11.24 Factor 2: correlation distance: (single)	352

11.25	Clustering Interval Time Series (ITS) on centers: long period of interval	353
11.26	Clustering Interval Time Series (ITS) on centers: short period of interval	353
11.27	Clustering Beanplot time series (BTS) using the model distance (method average)	354
11.28	Differenced time series from the Beanplot Clustering process	355
11.29	Interval attribute time series DJI (first 100 observations)	358
11.30	Boxplot attribute time series DJI 1900-2011	358
11.31	Forecasting Beanplot time series (BTS) using the mix- ture: coefficients estimation	361
11.32	Forecasting beanplot time series (BTS) using the mix- tures: factor time series	362

Introduction

Big data and huge or massive datasets are becoming ubiquitous. At the same time there is a growth of applications that collect data in real time, for example internet databases or financial data. So the general problem nowadays is the need to work extensively with huge data. In these cases, it is not always possible to store the data in such kinds of databases.

In all cases, data represents value that can be exploited through the extraction of the information contained within it for business aims. So the challenge for Data Science is to consider new methods to extract the information on huge datasets and to use it for the creation of value.

In this work the main focus is on high frequency data. These data rely on phenomena that generate unequally spaced observations, with the particular characteristics of an overwhelming number of observations over time, erroneous data, price discreteness, volatility, etc.

Recently, in literature new types of structured data have been proposed, which have an internal variability: intervals, boxplots and histograms. We introduce this structured data as representation in concrete problems and applications related to such kind of data.

In particular, the applicability of these representations to time series analysis of very long time series has been studied. In the most recent literature there has been the application of time series representations using the Histogram and Interval Data, these have been

applied to real problems like the analysis of financial time series, and there has been the analysis of time series related to other sectors like energy, etc.

A relevant problem is the temporal interval to choose in order to optimally define structured data. Various options are possible: hour, day, week, month, and year. A clear answer depends on the specific application we are interested in. So, sometimes, it may be useful to consider structured data, by considering the hour for trading applications, and sometimes it could be useful to consider a bigger temporal interval: for example in analyses useful for risk management. Therefore, a specific best temporal interval does not exist. The choice can also influence the methods used in the analysis, as we will see during the thesis. In the thesis we propose a new structured representation, based on special data as densities or beanplots: the density (or the beanplot) time series. The density time series (or the beanplot time series BTS) is particularly useful for exploratory data analysis of high frequency data in which we can discover important information that could otherwise be lost. This type of representation could be particularly fruitful when we have a higher number of high frequency observations for each structured data.

In the thesis, starting from this type of visualization we consider the need to model the data. The aim of the modelling is for both clustering and forecasting. In particular, we propose two types of approaches and we show the advantages and disadvantages. The proposed coefficients estimation and representation by descriptor points allow us to use some specific forecasting models and to analyze in depth the structural changes and the existence of groups of beanplots time series with similar characteristics over time.

In forecasting, the selection of the best information set available in the models is crucial. With regard to this there is the use of an algorithm to select the best information available in the past, which we use, and which can be applied to update the predictions. The the-

sis is accompanied by simulation exercises as well as by applications and examples based on real data. All methods proposed have been implemented using algorithms written in R code (shown in the Appendices).

The thesis is organized as follows:

Part I: The State of the Art

Chapter 1. We consider the basic problems with huge data and the evolution that they have undergone in recent years. We analyze the responses of methods for the analysis of this database. In particular, the analysis of data such as interval data to boxplots or histogram is considered as a possibility to account for large databases without the information loss due to the aggregation of the data. This chapter reviews various methods and techniques related to the internal representations and the symbolic data. Then, the methods for starting from a relatively huge data base leading to operational data bases with data that serve as internal representations are described. The final internal representation data can be modelled internally to obtain the data models. In any case these data are characterized by the intra-period variation.

Chapter 2. The database of the new type (as seen in Chapter 1), with specific reference to the financial sector, is analyzed. In particular, an innovation of recent years has been that of making use of high frequency data (High Frequency Data) that calls for specific techniques in their econometric and statistical treatment. In the same way, the characteristics of financial time series which require the same use of specific techniques for data analysis are analyzed.

Chapter 3. An interval is given as the first type of internal representation that can be considered and analyzed. In particular, these types of representations have an algebra that is reviewed as the basis of the techniques considered later in the thesis. The evolution of the techniques of interval data (which are compared with the techniques proposed later in the thesis) is then discussed. In this sense, the tech-

niques for analyzing time series data interval are considered.

In Chapter 4, representations based on different data types other than Intervals, for example boxplots, histograms and more recently candlestick charts, are considered. Here we consider the histogram algebra as an extension of the interval algebra. At the same time, in the chapter the developments in time series analysis of boxplots or histograms are considered. The final problem is the internal representation that could be used: in particular the choice of the representations in concrete problems. We propose some considerations on the choice of the statistical data, which are extensively considered during the second part of the work.

Chapter 5. A clear alternative to the use of data histogram are new types of data defined as data density that produce data using the methods of Kernel Density Estimation. In particular, such a methodology allows us to obtain a smoother image of the underlying data structure, by choosing appropriate bandwidth and kernel. Using density data offer some advantages with respect to intervals or histogram data in main precise applications: In particular, for large database this type of data can approximate in a better way histogram representation. The Kernel Density Estimation is analysed and its characteristics and properties evaluated. In particular, we focus on the choice of bandwidth and kernel.

Part II: New Developments and New Methods

Chapter 6. From the Kernel Density Estimation of Chapter 5 the data given in Beanplot density used in the thesis is introduced and defined. In particular, we introduce the time-series data density (or Beanplot). Then we analyze the ability to display and explore this data in comparison with objects of different types, in particular in different data structures. The chapter is accompanied by simulation exercises that consider simulated high frequency data comparing different types of time series of complex objects.

Chapter 7. In this chapter methods for analyzing time series of

Beanplot Data Models are considered. In particular, we describe two types of approaches: the first leading to a fundamental description of the dynamics of Beanplots over time, and the second, which separates the structural aspect of Beanplots compared to the "noise" (or the error). It is assumed that the data are characterized by patterns of Mixture Models. In this sense, the coefficients are used to capture the evolution of these models over time. The models, therefore, can highlight the change in inter -or intra-temporal of huge data and replace the series of Beanplots with coefficients (which are considered in their temporal evolution).

Chapter 8. At this point, there is the need to take into account the time series of attributes and trajectories obtained to build appropriate predictive models. In particular, the identification of the forecasting model to estimate each point of the series. In this case there can be the need to use different approaches in forecasting and combine the forecasts obtained in the procedure. The goal is to minimize the risk and uncertainty in the choice of a unique forecasting model in the presence of very volatile data, structural changes or model parameter changes (Parameter Drift). Finally, a search algorithm is applied to identify the range of observations to be used in the optimal forecasting model.

Chapter 9. We analyze the problems of Beanplot Clustering for time series. In particular, starting from the time series of attributes obtained we synthesize the beanplot time series (BTS) by using the Time Series Factor Analysis-TSFA, which synthesize the Beanplot dynamics over time.

For the Cluster Analysis, various types of Correlational Distances for Beanplot time series (BTS) have been considered, whereas suitable distance model (as proposed by Romano, Giordano and Lauro) have been used when the Beanplot are represented.

Chapter 10. In this chapter the performance of both internal and external models are analysed on the basis of indices of the adequacy of

the models. In particular, the evaluation and validation of models can lead to the respecification of both the internal models and the external models referred to in Chapters 7 (visualization and data exploration) 8 and 9 (regarding the identification and construction of the external models).

In Chapter 10 we consider real case studies in the field of Financial Market Monitoring, Asset Allocation, Statistical Arbitrage and Risk Management using the methods seen in the thesis.

In the final Chapter there are conclusions and future developments.

In the Appendices there are the R codes which replicate the procedures proposed and analyzed during the PhD thesis.

Part I

Data: The State of The Art

Chapter 1

The Analysis of Massive Data Sets

”From now on, the key is knowledge. The world is not becoming labor intensive, not material intensive, not energy intensive, but knowledge intensive” says Peter Drucker, the authoritative manager and consultant in 1992 [238]¹.

At the same time it was recently stated that ”The statistics profession has reached a tipping point. The need for valid statistical tools is greater than ever; data sets are massive, often measuring hundreds of thousands of measurements for a single subject. The field is ready for a revolution, one driven by clear, objective benchmarks by which tools can be evaluated²..” (Der Laan, Hsu and Peace 2010 [672]).

In this sense, the big data are becoming a growing flow in every area of the economy (McKinsey 2011 [499] and the Economist [655]

¹Bifet Kirby 2009 [80]

²In this sense there is the birth of ”Data Science”. See Loukides (2010) [468] for a detailed analysis about the reasons and the perspectives of the discipline. The author explains in the Report that ”the future belongs to the companies and people that turn data into products”

and Science 2011 [616]³. These data are typically timeliness, and in real time (Mason 2011 [490]) intrinsically a value⁴

In particular big data are datasets which grow in a way that are difficult to be managed using on-hand database management. Difficulties can be considered in a wide sense, for example in information capture, in data storage as in Kusnetzky 2010 [438], in the information extraction and search using adequate tools, in sharing information and reporting, in analytical methods (Vance 2010 [674]) and in the visualization of these data⁵ (Boyd and Crawford 2011 [107])

McKinsey 2011 [499] gives a definition of the relevance of the Big Data concept: "Big data refers to datasets whose size is beyond the ability of typical database software to capture, store, manage, and analyze"⁶. The definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data i.e. we do not define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes).

It is probable to assume that as technology advances⁷ over time then

³In particular the big datasets are interesting for the problems which they can solve in business and for the capability to create value, the concept is clear in Lev Ram (2011) [454] in which the author interviews Jim Goodnight, CEO of software maker SAS

⁴Madsen 2011 [474] "...In their data there is a competitive advantage"

⁵The evaluations on the problems of Big Data given, its enormous advantages are growing; see for example: The Economist 2010 [655], [2] and [730]

⁶Manovich 2011 observes: "There is little doubt that the quantities of data now available are indeed large, but thats not the most relevant characteristic of this new data ecosystem. Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.."

⁷See for example Miller 2010 [506]: "Cloud computing and open source software are fueling the data and analytics binge. The cloud allows businesses to lease

the size of datasets that qualify as big data will also increase⁸ Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a dozen terabytes to multiple petabythes (thousands of terabytes).

By considering this point it is necessary to stress the fact that the volume of data is growing at an exponential rate (see McKinsey 2011 [499]). There are in that sense various research works investigating this growth over time. Lyman and Varian, as reported by the McKinsey Report in 2011 [499], "estimated that the size of new data stored, doubled from 1999 to 2002 at a compound annual growth rate of 25 percent".

In that way, in recent years huge datasets have become ubiquitous because of the number of systems or applications which produce large volumes of data (see Aggarwal 2007 [7]). In particular during the past few years it has been very easy to collect huge amounts of data, also defined as "massive data-sets". Examples [576] (see Raykar

computing power when and as they need it, rather than purchase expensive infrastructure. And the combination of the R Project for Statistical Computing and the Apache Hadoop project that provides for reliable, scalable, distributed computing, enables networks of PCS to analyze volumes of data that in the past required supercomputers. With the Hadoop platform, Visa recently mined two years of data, over 73 billion transactions amounting to 36 terabytes. The processing time dropped from one month to 13 minutes"

⁸Where the size of data sets increase they become more and more real time, see Babcock [52] 2006: "But databases aren't just getting bigger. They're also becoming more real time. Wal-Mart Stores Inc. refreshes sales data hourly, adding a billion rows of data a day, allowing more complex searches. EBay Inc. lets insiders search auction data over short time periods to get deeper insight into what affects customer behavior". The problem is well known also for financial data in which econometric techniques to face high frequency data use some special techniques in real time, see: Pesaran and Timmermann 2004 [558]

2007) can include as the author did, genome sequencing, astronomical databases, internet databases, medical databases, financial records, weather reports, audio and video data. At the same time, it is possible to consider in practice other data typologies (see Huang Kecman Kopriva 2006 [372]).

So where modern database are diffused everywhere in industrial companies and public administration (Diday 2008 [209]) they tend to increase dramatically their size and the technical advances in databases and information systems are continuous (see for example the annual conferences organized in very large databases (Very Large Data Base Endowment Inc (2010) [683]) and the O'Reilly Strata Conference (2011) [541]).

At the same time the company IDC financed by the EDC has completed some research on the "Digital Universe" showing that the amount of digital data exceeded the world's data storage for the first time (cite Gantz et al. 2008 [287] but also Gantz and Reinsel [289] and [288]), where "the digital universe will be 10 the size it was 5 years before".

This result was very important because there are no possibilities to store the data created at all, and the rate of creation of the data generated grows to a higher level, thus exceeding the data storage capacity (See fig.1.1 and fig.1.2), so the gap between the two is continuously growing (see McKinsey 2011 [499]). Another work cited in McKinsey by Hilbert and Lopez 2011 [361] investigates storage capacity: global storage capacity grew annually at an annual rate of 23 percent over the period 1986-2007 whereas the data stored in digital form increased to 94 percent in 2007.

At the same time there are limits in the capability of processing this amount of data (see McKinsey 2011 [499]) when considering sensory and cognitive abilities. For example it was studied that the brain in its short-term memory can handle seven pieces of information (see Miller 1956 [505]). So another important problem considered was in-

formation overload⁹.

To solve these problems there are various possible solutions, as for example, using more sophisticated methods or algorithms or using different types of data that could be used, studied extensively during this thesis (Schweizer 1984: "Distributions are the Numbers of the Future" [615]).

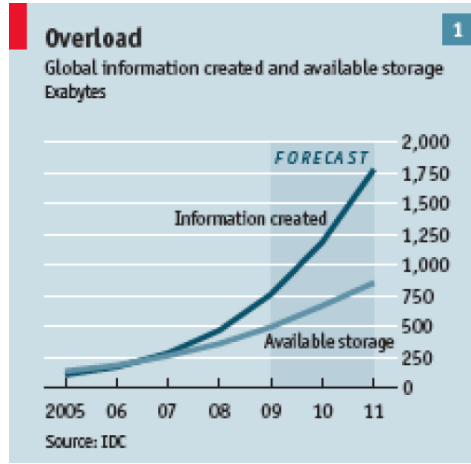


Figure 1.1: Global Information created and available storage 2005-2011. See The Economist 2010 [655] and Batini 2010 [65]

1.1 Complex Data Sets and Massive Data

Let a data matrix $H_{n,m}$ be an $n \times m$ observation \times variables, where $w_{n,m}$ are scalar data, so we have:

⁹See also Makarenko 2011 [478]

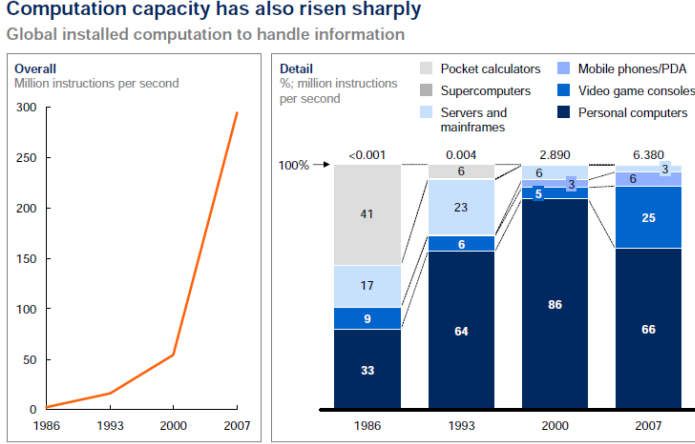


Figure 1.2: Computation Capacity 1986-2007 see Hilbert and Lopez 2011 [361] and McKinsey 2011 [499]

$$H_{n,m} = \begin{matrix} & m_1 & m_2 & \cdots & m_m \\ \begin{matrix} n_1, \\ n_2, \\ \vdots \\ n_n, \end{matrix} & \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,m} \end{pmatrix} \end{matrix} \quad (1.1)$$

In this respect, complex data can be defined: "Any data which cannot be considered as a standard observation \times standard variables data table" (Diday 2011 [214]). It can be considered Complex Data: several data tables describing different typologies of observations. Specific examples can be considered in various works (for example Diday 2011 [214]):

1. Hierarchical Data

2. Textual Data
3. Time Series Data in each cell
4. Multisource Data Tables (Data Fusion)

Massive Data, are datasets of huge dimensions, and they come from many sources¹⁰, and they can be generated by various devices like sensors, cameras, microphones, pieces of software (see Huang Kecman Kopriya 2006 [372] and Diday 2008 [209]). Another important domain, in the enormous increase of the data, can be considered due to Data Streams Applications (Aggarwal 2007 [7] and Balzanella Irpino Verde 2010 [56]).

A typical example of the differences of the data stream applications with respect to the data mining procedures is given by Domingos Hulten 2010 [224], data, in particular are collected, in various applications faster than it is possible to mine (Balzanella Irpino Verde 2010 [56]). In this respect, to avoid data losses it is necessary to pass to systems that are able to mine continuously the high-volume, open-ended data streams at the time they are available.

Clearly in other situations, also with huge data sets it is possible to use the classical data mining approach. Data are everywhere so the approach can be generalized to different fields.

A typical example of complex data are also High Frequency financial data (See Dacorogna et al. 2001) [163]. In finance the innovation was typically due to the introduction of Tick Data (also defined High Frequency Data) that made it possible to develop trading strategies taking into account Intraday market movements. The empirical study of these dynamics would be very beneficial for an understanding of the

¹⁰In particular the relevant information comes from many data sources where the problem, today, becomes how to combine this information (Ras Tsumoto Zighed 2005 [577])

markets and the reduction of the associated risk of the price fluctuations (see Engle Russell 2009 [254]). High Frequency Data in particular can help to forecast risk (see for example Kaminska 2008 [415]).

At the same time complex time series can be also obtained by considering some time series in large data sets (for example in Finance or in Energy applications like Load Forecasting) where the series are specifically characterized by lower frequency but at the same time by "complex" characteristics (spikes, nonlinearities, high volatility etc.). In this sense it is possible to consider the approach of the time series as complex data (Diday 2008 [209]).

It is important to stress the fact that using large datasets is a specific need (and sampling in that case does not help), because the real data are huge and continually flowing, but at the same time it is a specific advantage (the creation of the value).

There are cases in which using big datasets can be very useful and this is typical for exploratory studies where in that case it is not sufficient to define some statistical relationships that could be adequately estimated or tested (see Benzecrí 1973 [74] Lebart Morineau Piron 1995 [452] Saporta 1990 [607] Gherghi Lauro 2002 [300] and Bolasco 1999 [100]).

At the same time it is possible to consider some other cases in which data are overwhelming and theoretical models need to adequately face up to the existing data (see Sanchez Ubeda 1999 [606]). In these cases big datasets can be useful to generate some hypothesis with data (Tukey 1977 [670]).

The general case and the most classical case in using big databases is Data Mining or Business Intelligence (Giudici 2006 [312]). In this case large datasets can be used to extract information and to extract the knowledge present in data. Here, the idea is that of using this specific knowledge to obtain some relevant business indications. In

this case the data are considered a specific richness to use¹¹.

The first and direct consequence of the data, in this sense is that humans cannot handle and manage such a massive quantity of data, which are usually collected in the numeric shape as the huge rectangular or squared data matrices (see Huan Kecman Kopriva 2006) [372]. The challenge in this sense is using specific systems that automatically extract the information from the raw data to permit better decisions (see Raykar 2007 [576]).

In this case it is necessary to apply some specific statistical techniques in order to achieve the data management and the knowledge extraction, that is, therefore we need to use specific statistical techniques to handle these data sets.

On the contrary, using the single valued variables (using for example some form of data aggregation) brings information loss. In the graph there is an example in which a large data set does not allow the observation of the data structure of the underlying financial data (See fig.1.3).

1.1.1 **Characteristics of Complex Data Sets and Massive Data**

Massive data are characterized by an overwhelming number of observations and/or variables. Problems in these datasets are related to (see McKinsey 2011 [499]):

1. Data Storage
2. Data Search and Extraction

¹¹Data richness allows the improvement of data analysis to a certain extent in order to improve the data analysis, see for example: Linoff Berry 2011 [461] and Weiss Indurkha 1996 [696]

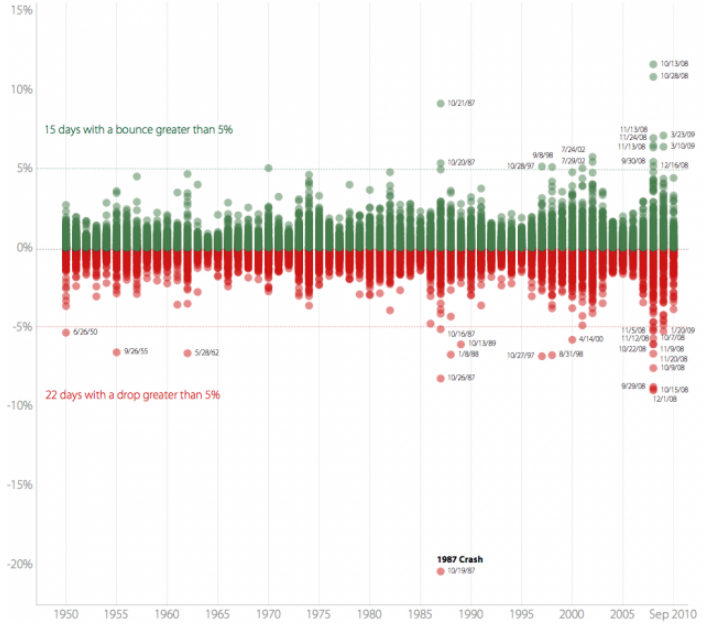


Figure 1.3: Daily Price Change of S&P Grouped By Year, data from Yahoo Finance. Source: VisualizingEconomics.com

3. Data Sharing

4. Analytics¹²

Generally in these types of datasets some data strategies are necessary to handle the data. These data strategies can be, for example, partitioning the dataset or aggregating the observations where necessary. Clearly these methods are continuously updated and they are under the continuous scrutiny of researchers. So there are various

¹²A problem related to Analytics is scalability (Berthold and Hand 2003) [76]

techniques and algorithms that could be used in these cases¹³.

Diday 2011 [214] states that a solution could be possible in differentiating standard scalar observations (classical data) from the symbolic observations (data that represents an internal structured variation and which are structured). In that sense we can have: "Standard observations like a player, a fund, a stock... Symbolic observations: Classes: a player subset, a subset of funds, stocks... Categories: American funds, European funds, ... Concepts: an intent: volatile American funds, an extent: the volatile American funds of a given data base."

There are important cases in which it is particularly useful to use the concepts instead of classical data, cases for example in which we are considering data where in itself the concept could be important, and cases in which we need to manage a data fusion of different data tables or datasets.

In particular what are the advantages of using Internal Representations or Symbolic Data? Diday 2011 [214] states them to be these:

1. Considering the right generality level of a collective data without information loss.
2. Reducing the data set size and so reducing the number of variables and observations (reducing computational costs of the analyses).
3. Mitigating the problem of missing data.
4. Ability to "extract simplified knowledge and decision from complex data".
5. Solving the problems related to confidentiality.
6. "Facilitate interpretation of results": decision trees, factorial analysis, new graphic kinds.

¹³McKinsey 2001 [499]

7. "Extent Data Mining and Statistics to new kinds of data with many industrial applications".

There are some cases in which data sets are characterized by many outliers or missing data. So it could be important to provide a data imputation (in the case of missing data) to allow a safe use of the aggregate representations. In fact, sometimes missing values are not distributed at random during the dataset and they are missing following a pattern.

In the case of missing data (that could be considered not a random), they need to be substituted using some statistical methods¹⁴. Various strategies could be considered for the original missing data: see Little and Rubin 1987 [464], Allison 2001 [15] and Howell 2007 [367].

In any case, a preliminary analysis on the data to detect the outliers and an imputation strategy (if there are missing data) is necessary. In fact both outliers and missing data can affect the statistical analysis.

1.1.2 Statistical Methods, Strategies and Algorithms for Massive Data Sets

Later we will analyse in depth statistical methods that consider data as representations (interval, histograms etc.). Many different approaches are considered in literature that could be used considering scalar data in massive data sets¹⁵. It is important to note that we can aggregate or not the entire dataset. In particular one possibility is to work on the entire dataset without any type of aggregation.

Various strategies and methods can be considered (see Giudici 2006

¹⁴Zuccolotto 2011 [725] for an approach on symbolic data

¹⁵A review and a presentation of some approaches is in Rajaraman, A. and Ullman D.J. (2010) [572] Gaber, Zaslavsky and Krishnaswamy (2005) [285]

[312]), alternatively we can consider many methods together (strategies) as for example in Gherghi Lauro 2002 [300] and Bolasco 1999 [100] in which we use more methods sequentially. So in these cases we can define different strategies for the analysis in which, for example, we reduce a dataset using a factorial method and after classify the statistical units¹⁶.

Relevant methods used to analyze Big Data in Businesses today are enumerated in the McKinsey 2011 [499] Report. This point needs to remain open because the approaches and advances in literature and in business evolve very quickly.

1.2 Analysing data using Aggregate Representations

A different approach is related to that considering Aggregate Representations (the entire representation expressing variation for the data disaggregated) and working with methods like those in Symbolic Data Analysis (see for example Diday 2008 [209] where Valova and Noirhomme Fraiture consider explicitly the case of massive data and Symbolic Data Analysis [673]).

In these cases we directly consider some types of new data, as for example intervals, histograms etc. These new data are structured and express internal variation on the single data. In particular the data can be defined symbolic data if they contain more complex information than scalar data (they can be characterized by internal variation and could be structured Diday 2002 [207]).

In that sense the symbolic data can summarize massive data bases

¹⁶It is possible to consider the approach of data analysis in SPAD software for example in which we perform the statistical and data mining analysis using "chains" or sequences of different methodologies see: Coheris 2011 [737]

by considering their Concepts. They can be defined as first units and be characterized by a specific set of properties defined as "intent" and "extent" that could be seen by the set of the units which suits these properties. The Concept could be described by symbolic data which can be intervals, histograms, etc.

The characteristics of the data allow us to bear in mind the internal variation of the "extent" by considering the different Concepts (Diday 2002 [207] and also 2008 [210]).

There are important cases in which Concepts are relevant and they were described by Diday in 2011 [214]. Symbolic Objects can be relevant in modelling the Concept as shown in Lauro and Verde 2009 [449]:

1. When there is a specific interest on the Concepts (for example when the data analysis is based on the respect to the single units)
2. "When the categories of the class variable to explain are considered as new units and described by explanatory symbolic variables"
3. In the case of data fusion of multisource tables

Another important preliminary analysis is the optimality of the Concept chosen¹⁷

1.2.1 Scalar Data and their Aggregate Representation

So we can specifically define the complex data as, for example, the Interval and the Histogram Value Data: complex data are data that cannot be considered standard observation \times variables or $n \times m$ data

¹⁷Diday 2011 [214]

1.2. Analysing data using Aggregate Representations

tables, interval and histogram data are typically data in which there exists a variation inside the classes of standard observation (see also Diday 2010 [211]). In that case, by starting from the initial massive data sets, each cell of the data table can contain an interval, a boxplot, an histogram, a bar chart, a distribution etc.

So we have different real valued vectors (see Billard Diday 2006 [87] and 2003 [85] and Signoriello 2008 [630]) in which n statistical units are evaluated by m variables, so a data table is a $n \times m$. Data on G random variables can be represented by a single point in a g -dimensional space \mathbb{R}^g . For example a classical data value x is a single point in a g -dimensional space $x = 12$. An example of classical data is provided in Billard 2010 [82], where she considers a medical data set (see See fig.1.4):

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
1	Boston	M	24	M	S	2	2	0	165
2	Boston	M	56	M	M	1	2	2	186
3	Chicago	D	48	M	M	1	3	2	175
4	El Paso	M	47	F	M	0	1	1	141
5	Byron	D	79	F	M	0	3	4	152
6	Concord	M	12	M	S	2	1	0	73
7	Atlanta	M	67	F	M	1	6	0	166
8	Boston	O	73	F	M	0	2	4	164
9	Lindfield	D	29	M	M	2	0	2	227
10	Lindfield	D	44	M	M	1	3	3	216
11	Boston	D	54	M	S	1	5	0	213
12	Chicago	M	12	F	S	2	2	0	75
13	Macon	M	73	F	M	0	3	1	152
14	Boston	D	48	M	M	0	2	4	206
15	Peoria	O	79	F	M	0	3	3	153

Figure 1.4: Classical data in a medical data set [82]

Interval data on G random variables can be g -dimensional hyper-cubes or hyperrectangles or a Cartesian product of g distributions¹⁸.

¹⁸Billard 2006 [81]

That is, by considering the case of the intervals $[\underline{x}_1, \overline{x}_1]$ and also $[\underline{x}_2, \overline{x}_2]$ where \underline{x} is the interval lower bound and \overline{x} the interval upper bound, with $g = 2$ where the random variables take values over the intervals. The data value in this case is the rectangle $R = [\underline{x}_1, \overline{x}_1] \times [\underline{x}_2, \overline{x}_2]$ and vertices of R are: $(\underline{x}_1, \overline{x}_1), (\overline{x}_1, \overline{x}_2), (x_2, x_1)$ and (x_2, \overline{x}_2) .

The $g = 2$ dimensional hypercube is a space in the plane. Where $\underline{x}_1 = \overline{x}_1 = \underline{x}_2 = \overline{x}_2$ we obtain a single point (like a hypercube of $g = 0$ and a special case). In that sense, classical data are a special case of symbolic data in which the point value is $[\underline{x}, \underline{x}]$ or also $[\overline{x}, \overline{x}]$. An example of interval data is provided by Billard [82] related to a Mushrooms data set (See fig.1.5):

ω_i	Species	Pileus Cap Width	Stipe Length	Stipe Thickness	Edibility
ω_1	<i>arorae</i>	[3.0, 8.0]	[4.0, 9.0]	[0.50, 2.50]	U
ω_2	<i>arvenis</i>	[6.0, 21.0]	[4.0, 14.0]	[1.00, 3.50]	Y
ω_3	<i>benesi</i>	[4.0, 8.0]	[5.0, 11.0]	[1.00, 2.00]	Y
ω_4	<i>bernardii</i>	[7.0, 6.0]	[4.0, 7.0]	[3.00, 4.50]	Y
ω_5	<i>bisporus</i>	[5.0, 12.0]	[2.0, 5.0]	[1.50, 2.50]	Y
ω_6	<i>bitorquis</i>	[5.0, 15.0]	[4.0, 10.0]	[2.00, 4.00]	Y
ω_7	<i>californinus</i>	[4.0, 11.0]	[3.0, 7.0]	[0.40, 1.00]	T
ω_8	<i>campestris</i>	[5.0, 10.0]	[3.0, 6.0]	[1.00, 2.00]	Y
ω_9	<i>comtulus</i>	[2.5, 4.0]	[3.0, 5.0]	[0.40, 0.70]	Y
...

Figure 1.5: Interval data in a mushrooms data set [82]

Different data can be lists $\{good, fair\}$ by considering one or more value in the list.

Classical single value data have no internal variation and usually they are forced to be a single value in large data sets (causing loss of information), where, interval and histogram data are characterized by internal variation: in particular by considering a value of g as $[\underline{x}, \overline{x}]$, with $\underline{x} \neq \overline{x}$ that could be considered as taking a continuum of values in the interval.

1.2. Analysing data using Aggregate Representations

There are other important cases in which the use of interval and histogram data are relevant (Signoriello 2008 [630]). It is for example the case of more complex information in data which call for more flexible representation. In this sense, the exact information of interest is not a real value but can be chosen from sets, intervals, histograms, boxplots, trees, graphs or functions. Here is an example in which the data of interest can be considered histogram data (See fig.1.6):

ω_i	Concept		Frequency Histogram
	Gender	Age	
ω_1	Female	20s	{[80, 100), .025; [100, 120), .075; [120, 135), .175; [135, 150), .250; [150, 165), .200; [165, 180), .162; [180, 200), .088; [200, 240), .025}
ω_2	Female	30s	{[80, 100), .013; [100, 120), .088; [120, 135), .154; [135, 150), .253; [150, 165), .210; [165, 180), .177; [180, 195), .066; [195, 210), .026; [210, 240), .013}
ω_3	Female	40s	{[95, 110), .012; [110, 125), .029; [125, 140), .113; [140, 155), .206; [155, 170), .235; [170, 185), .186; [185, 200), .148; [200, 215), .043; [215, 230), .020; [230, 245), .008}
ω_4	Female	50s	{[105, 120), .009; [120, 135), .026; [135, 150), .046; [150, 165), .105; [165, 180), .199; [180, 195), .248; [195, 210), .199; [210, 225), .100; [225, 240), .045; [240, 260), .023}
ω_5	Female	50s	{[115, 140), .012; [140, 160), .069; [160, 180), .206; [180, 200), .300; [200, 220), .255; [220, 240), .146; [240, 260), .012}
ω_6	Female	70s	{[120, 140), .017; [140, 160), .083; [160, 180), .206; [180, 200), .294; }
...
ω_{14}	Male	80+	{[155, 170), .067; [170, 185), .133; [185, 200), .200; [200, 215), .267; [215, 230), .200; [230, 245), .067; [245, 260), .066}

Figure 1.6: Histogram data in a Cholesterol data set Gender \times Age categories: (Billard 2010 [82])

The most important difference between interval and histogram data with respect to scalar data is that the first one shows an internal variation.

Considering a specific data set, we can identify some regrouping criteria (in a credit card data set for example, the specific transactions for each person over time) and in that way can define accordingly the data summary. Various possible summaries need to be considered.

In each of these examples the data can be transformed into single valued data but interval data shows a higher complexity. Billard and Diday 2010 [88] in this sense make various examples: transactions by dollars spent $[5, 1200]$ or a different summary by type of purchase (gas, clothes, food, ...) or, by type and expenditure (gas $[30, 60]$), food, $[25, 105]$...). In all these examples it is necessary to consider as well a temporal component as the summarized values over the time t , for example transaction by dollar in various periods.

Interval and Histogram data can capture specific variation over time t , important in their own right on any data set. Variables g can be collected as an interval over time t (for example [88]: pulse rate at time t ($[60, 72]$), at time $t + 1$ ($[62, 74]$) systolic blood pressure at t ($[120, 130]$), at $t + 1$ became ($[122, 132]$) and diastolic blood pressure ($[85, 90]$) at t and ($[87, 93]$) for each of $n = 100$ patients (or, for $n = 12$ million patients). Alternatively, we can consider the evolution over the time t of different classes of $n = 31$ students that could be characterized by boxplots, histograms or distributions of their marks for each of several variables g at time t (for example mathematics, statistics and biology).

The information loss in aggregate data is shown by Billard 2006 in [81]. In fact it is possible to show that, considering three realizations of a random variable $G = \text{weight}$ (accordingly with the considered dataset), we have $G_1 = 135, G_2 = [132, 138], G_3 = [129, 141]$. It is possible to consider these three samples each of size $n = 1$. Bertrand and Goupil 2000 [77] show that S^2 is specifically the sample variance of each variable G under the assumption of uniformly distributed values in each interval, where: $P_u = [\underline{x}_u, \bar{x}_u], u = 1 \dots n$. We have:

$$S^2 = \frac{1}{3n} \sum_{u=1}^n (\underline{x}_u^2 + \underline{x}_u \cdot \bar{x}_u + \bar{x}_u^2) - \frac{1}{4n^2} \sum_{u=1}^n [\underline{x}_u + \bar{x}_u]^2 \quad (1.2)$$

So we obtain that the sample mean is $\bar{P}_1 = \bar{P}_2 = \bar{P}_3$ where

$S_1^2 = 0$, $S_2^2 = 3$, $S_3^2 = 12$. For the basic statistics procedure in Interval Data Analysis see also Gioia Lauro 2005 [310] and Billard [82]. The internal variation of the interval and histogram observation determines the difference between the three results. In this case we can show that it is necessary to take into account the internal variation considering interval and histogram data.

It is important to note that in all these cases these types of data are inherently rich in nature, infact in these cases data are characterized and can be compared by not only a single value, but at the same time by a location (say, the central value), a size (the internal variation) and a shape (the exact form of the distribution).

It is possible to find a specific link between data collection and its representation (for example, after a query in a database) it is possible to find the same link between the interval data and its interpretation. Data that show relevant internal variation (due to the data internal heterogeneity) need to be analysed using specific statistical techniques (interval and histogram valued data analysis techniques in particular).

Interval and Histogram value data arise in different ways. In fact the data we have considered are natively interval and histogram (they show an internal variation that could not be represented as a scalar data). So the single values do not represent faithfully the data we want to consider. If a data is natively an interval and we force the data to be a single-valued data we are forcing the data to be scalar and we are not considering its real nature of interval. In this sense the use of the interval data is determined by the nature of the original data

In other cases we can be specifically interested not in the single value but in the specific variation, because the single value might not be so relevant (due for example to fluctuations of the measurements). In that sense there are real cases in which it is very difficult to measure a specific phenomenon as a single scalar or value (due to a specific reason) and in that case single observations would not be relevant. In

this case an interval or a symbolic data can capture in a better manner a real phenomenon by considering the intervals of the measurements.

Fluctuations can be related specifically to errors both in the data and in the solutions. Gioia and Lauro 2005 [310] provide some examples of these errors:

1. Measurement errors: where the measured value of a physical quantity x may be different from the exact value of the quantity
2. Computation errors: when round errors make a distortion from the true results due to the finite precision of the computers
3. Errors due to uncertainty in the data: the value of a specific data cannot be measured precisely in a physical way

Billard and Diday 2010 [88] show other cases in which there are no relevant errors in data, but actual technologies do not allow the performance of the requested computations.

So we can have a fourth case in which the use of symbolic data is important. By considering n observations (when n is very large with hundred of thousands or more) with m variables (at the same time with m hundred or more), so by taking into account a $n \times m$ matrix H in an inversion computation H^{-1} the computational burden can be relevant.

It is important to note that also where computer capabilities expand at the same time (larger computation of H^{-1} at a time, the burden assuming a growth either of n and m will be relevant) it is important to consider the growth of the dataset size and so the size of the $n \times m$ matrix H (Gantz et al. 2008 [287]).

The last reason for the use of symbolic data is when there are problems with results, where there is aggregated data that does not faithfully confirm the results obtained by disaggregated data. The result is well known for example in High Frequency Data in which there

can be explicit differences in analyses considering different types of aggregations, for example, frequencies (see Dacorogna et al. 2001 [163]). Differences between results using aggregated and disaggregated data are well known as well in Econometric literature (aggregation bias). For example in the presence of outliers it can be very dangerous to use data aggregations because the methods used may not be robust (in a statistical sense) and results used may not faithfully represent the "real" data.

At the same time, aggregating data presenting outliers can lead to the loss of the information related to the original problem. In this sense outliers cannot be detected in the aggregated data, or they can be masked by the data structure.

Infact, it is important to note that aggregation cannot capture the real structure of the data but tends to force data to have a single value. Also by considering more complex aggregation methods, for example robust methods, we lose information. For this reason, Schweizer 1984 [615] says that "the distributions are the number of the future!", thus we need to consider statistical approaches that directly use distribution data. In these cases we need to consider these data as internal representations or symbolic data.

1.2.2 **Sources for Aggregate Representations and Symbolic Data**

Symbolic data can be obtained in the process of data reduction from a huge dataset. Any query in a database can produce descriptive variables and categories. Diday [207] shows, for example, categories (SPC) crossed with categories of age (Age) and regions (Reg). So it is possible to obtain a new categorical variable of cardinality $|SPC| \times |Age| \times |Reg|$ where $|X|$ is the cardinality of X . Another important

way to obtain Symbolic data is by considering a clustering process (see also Diday 1993 [203] for the steps of a Symbolic Data Analysis). In this sense we obtain a Symbolic data naturally from the classes obtained.

Diday [207] also states that Symbolic data can be at the same time "native". If for example, they become:

1. Expert Knowledge
2. Any random variable ("from the probability distribution, the percentiles or the range of any random variable associated to each cell of a stochastic data table")
3. Time Series
4. Confidential Data
5. Relational Databases (merging different relations in order to analyse a set of observations)

Describing the process from the relational databases to symbolic data is the objective of Hebrail and Lechevallier [641]. In particular, in this article, there is the two level paradigm where a symbolic object could be created by considering a process of aggregation of single individuals. The authors describe the generalization process of a classical dataset extracted from a relational database.

1.2.3 Complex Data and Tables of Aggregate Representations

Following Diday 2008 [209] the process of setting a symbolic description of the set of the individuals (or statistical units) is called the generalization process.

1.2. Analysing data using Aggregate Representations

Consider the concept of "swallow", for example, that could be characterized at time t by a description vector d :

$$w_{n,m} = (["yes"], [60, 85], [90\% \text{ yes}, 10\% \text{ no}]). \quad (1.3)$$

The values evolve at time t . The generalization process is relevant because it takes specifically into account the internal variation of the description of the individuals (a group of companies, for example) inside the set of individuals. For example, a set of swallows at time t on the island vary in size [55, 82]. In financial markets this variation in a portfolio can be considered a risk indicator.

It is important to consider that the initial form of the complex data sets in these types of matrices can also be reduced to aggregated data tables by means of adequate queries. This is the simplest way to obtain a data set of aggregate representations and symbolic data (figure 1.7 and figure 1.8 represent the steps from a complex data table to a symbolic or aggregate data table).

In the temporal context it is possible to decide different temporal intervals to obtain the data (for example, a typical high frequency time series could be aggregated by considering hours or days or it is possible to aggregate different series of a portfolio).

At the same time it is usually necessary to pre-process data in order to handle missing values and outliers accordingly (in the high frequency context, for example, there is a problem in this sense¹⁹). Besides the problem of data pre-processing, there is the need to transform complex data tables into symbolic data tables in each specific environment. There is a particular exception: that is the case in which the stream of the data arrives quickly and it is not possible to make any intervention.

In practice, a symbolic data table can be considered a table in which the columns are symbolic variables used to describe a set of units, de-

¹⁹Brownlees Gallo (2006) [115]

defined as the individuals considered or the statistical units (see Tzitzikas 2004 [671]). In these data tables the rows can be symbolic descriptions of the individuals.

At the same time, in symbolic time series data tables, in the rows there are the realizations of the series by considering the different interval temporal. In the case of the aggregate representations or symbolic time series data tables each column can be related to a different time series (for example to different stocks in a portfolio or different sensors etc.). It is possible to see the differences in symbolic data tables in Diday 2010 [211].

At this point, classical data tables are datasets with classical data, whereas symbolic tables are tables in which for each cell there is specific symbolic data, such as an interval, a histogram etc.

Internal representations and symbolic data tables can be considered in the same way as classical data tables but they are extensions of the classical data tables, where they can contain both classical data and internal representations.

A classical data table can be transformed into a symbolic data table while the contrary is not possible, because of the information loss (see Diday 2010 [211]). In fact, when symbolic data are transformed into classical data there is a data aggregation process, therefore the variation of the data is definitively lost (see Billard 2006 [81] and Bertrand Goupil 2000 [77]). In analyses the internal variation of the complex data is important, thus it is relevant to represent this variability.

It is important to note that often the internal representations in tables show characteristics of the original data that suggest the need to handle these data with symbolic data analysis methods. In fact, aggregate or internal representation can show some features of the data that allow us to extract the underlying "signal" by extracting the "noise" from the original complex data. This is the case, for example of particular nonlinear relationships in data.

An important problem is the identification of the subsets of differ-

ent distributions in data and at the same time the outliers, in fact, in this way there is a specific problem of overgeneralization. Diday states that overgeneralization, for example, happens when smaller and greater values characterize a numerical variable that generalizes the interval (see Diday 2008 [209]).

In particular, in choosing the optimal internal representation or symbolic data, it is important to identify if the overgeneralization phenomenon is relevant in the original complex data by detecting the outliers. At the same time, it is very important to identify mixtures in data that could be relevant as insights in data analysis. So there is a specific need for statistical methods that could identify such data features as outliers and mixtures (see Atkinson Riani and Cerioli 2004 [47]).

To detect outliers it is necessary to implement some outlier tests previously in the data pre-processing, for example the Grubbs Test (Grubbs 1969 both the studies [327] and [328]).

The problem of overgeneralization is at the same time related to the problem of choosing the optimal Concept and the optimal Symbolic Data. In choosing the optimal Concept various procedures can be used in the analysis (Diday 2011 [214]):

1. By considering the Hierarchical Data.
2. By clustering and optimizing the number of clusters considering various criterion such as AIC Akaike Information Criterion , BIC Bayesian Information Criterion.
3. By considering several clustering methods.
4. With the optimization of the discrimination between the Concepts and by considering the informative power of the histograms or bar charts.

Patient	Hospital	Age	Smoker
Patient 1	Hospital 1	74	heavy
Patient 2	Hospital 1	78	light
Patient 3	Hospital 2	69	no
Patient 4	Hospital 2	73	heavy
Patient 5	Hospital 2	80	light
Patient 6	Hospital 1	70	heavy
Patient 7	Hospital 1	82	heavy
Patient 8	Hospital 3	74	heavy
\vdots	\vdots	\vdots	\vdots

Hospital	Age	Smoker
Hospital 1	[70, 82]	{light 1/4, heavy 3/4}
Hospital 2	[69, 80]	{no, light, heavy}
Hospital 3	[74, 74]	{heavy}
\vdots	\vdots	\vdots

Figure 1.7: The process of transformation from complex data tables into symbolic data or aggregate representation tables [81]

1.2. *Analysing data using Aggregate Representations*

In this thesis we will consider specifically those tables in which data are both related to different time series or in which statistical units are considered and followed over the time.

In particular, classical tables in this sense are longitudinal or cross sectional data. So we consider data that represent single groups of time series (e.g. financial portfolios) because we wish to examine the differences in cross-sectional behavior or in its evolution. An example of these data is presented in Diday 2008. [211]

Whilst the transformation into symbolic data table is possible by considering classical data in symbolic data, at the same time it is also possible to consider transformations by other types of data, for example fuzzy data, into internal representations or symbolic data (see Diday 2008 [209]).

Multisource data tables arise specifically from the union of various data tables in which they are unified in a specific query. In effect, data are derived from various types of symbolic data arising from different data queries.

The aim of the symbolic data analysis is that of analysing symbolic data tables that describe observations with a variation in their description (Diday 2008 [209]). There are four types of analysis in internal representation and symbolic data analysis that we can consider (table 1.1):

Table 1.1: Data Analysis Typologies

Method	Classical Analysis	Symbolic Analysis
Classical data	Case I	Case II
Symbolic data	Case III	Case IV

In particular, in Case I we refer explicitly to the classical analysis of time series, Case II consists of the specific extraction of the symbolic

descriptions from classical datasets, whereas in Case III symbolic data are aggregated and transformed into Classical data, and in Case IV it is specifically the case of Symbolic Data Analysis (see Diday 2008 [209]).

In this work we will obtain the data as symbolic data and analyse these data using the internal representation and a symbolic data analysis. In this sense, the present research has produced various statistical methods that could be used in various contexts and situations²⁰

1.3 Aggregate Representations from Time Series

By considering the case of the time series, various proposals have been made: in particular Ferraris Gettler Summa Pardoux and Tong 1995 [268] Gettler Summa and Pardoux 2000 [297] and more recently Gettler Summa et al. 2006 [298] and Gettler Summa, Goldfarb [296] .

The specific steps in an analysis of temporal symbolic data are represented in figure 1.8 and are described in detail in Diday 2008 [209]. In particular, for the different stages of a Symbolic Data Analysis see Diday 1993 [203]:

1. The symbolic data analysis needs to be conducted on two levels: the first level is related to original observations during time t , on the second level are the concepts of the considered temporal intervals.

²⁰For an extensive review of statistical and quantitative methods in Symbolic Data Analysis see Diday 2008 [209] and 2011 [214], Diday and Noirhomme 2008 [218]. Actual software developments in Symbolic Data Analysis are represented by the Sodas software [743] (see Diday and Noirhomme 2008 [218]) and the recent Syrokko [742]. Various packages in Symbolic Data Analysis are proposed in R.

1.3. Aggregate Representations from Time Series

2. Each symbolic description describes the concept.
3. The description needs to take into account the variation.
4. Symbolic data analysis can extend the standard analyses to the cases in which observations are represented by symbolic data.

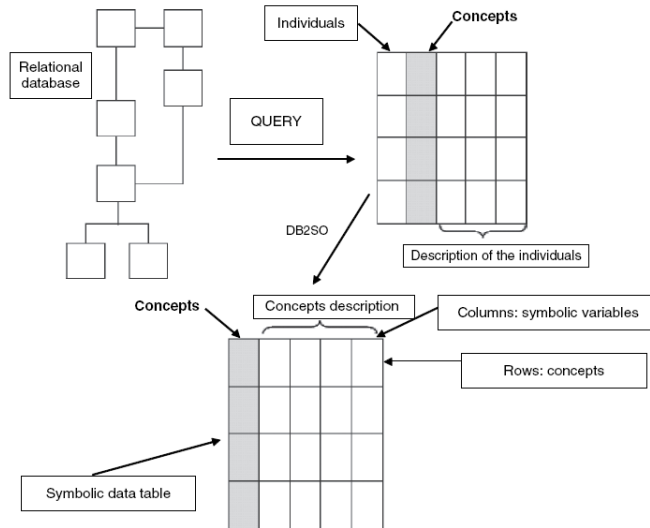


Figure 1.8: The process of transformation from a relational data table into a symbolic data table[209]

In this respect to apply specifically the tools of statistics to symbolic data, new statistical tools need to be considered those which take into account the characteristics of the symbolic data. Typically a Symbolic Data Analysis can be characterised by various phases (see Diday 2008 [209]).

Clearly the case here is related to a dataset H suggesting that these

data should be handled statistically. In other cases, for example in the data stream, we consider only the symbolic data as histogram data because we cannot handle the single observations directly (in this case data cannot be collected at all: Balzanella Irpino Verde 2010 [56]). We follow this approach:

1. It is assumed to start by considering more than one time series in a specifically built relational database. In this case the classical data set is obtained.
2. The series are pre-processed to detect and manage outliers.
3. The series are pre-processed in order to take into account the missing values and the imputation.
4. The specific interval is chosen (by defining a specific query).
5. Overgeneralization is checked (the appropriate symbolic data to use and their appropriate structure are chosen).
6. The symbolic data table is obtained.

1.4 A study simulation on Big Data and Information Loss

Here we present a simulation study²¹ considering big datasets and information loss in the aggregation process. In particular, we will simulate various time series models based on a large quantity of data, and we aggregate these data using various aggregation functions (the mean and/or the median) for a limited number of periods. At the same time we compute the interval of the minimum and the maximum in

²¹See Koenker 1996 [429] for the design of a simulation study

each period. The aggregation is necessary because of the difficulty of visualizing the original time series (we will consider this point in depth in chapter 6). The interval shows clearly the information losses of each subperiod of i observations. An indicator of information loss could be:

$$IL = \sum_{i=1}^n |\overline{x_i} - \underline{x_i}| \quad (1.4)$$

To compare the results between different periods and different aggregations some t -test are used. These confirm the result expected. In particular: the higher the aggregation interval considered, the higher is the information loss.

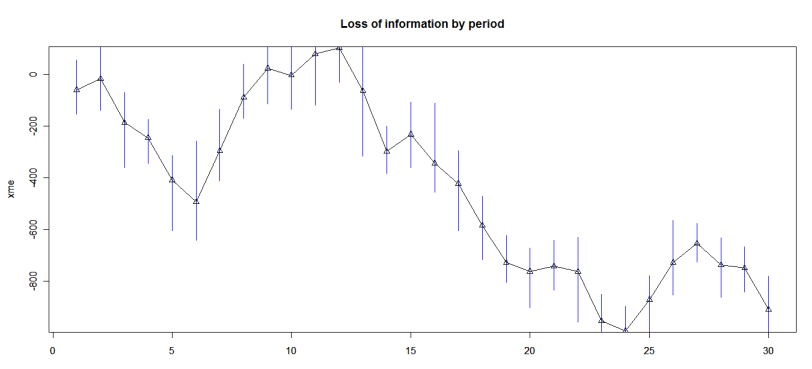
1.5 Applications on Real Data

1.5.1 The Symbolic Factorial Conjoint Analysis for the Evaluation of the Public Goods

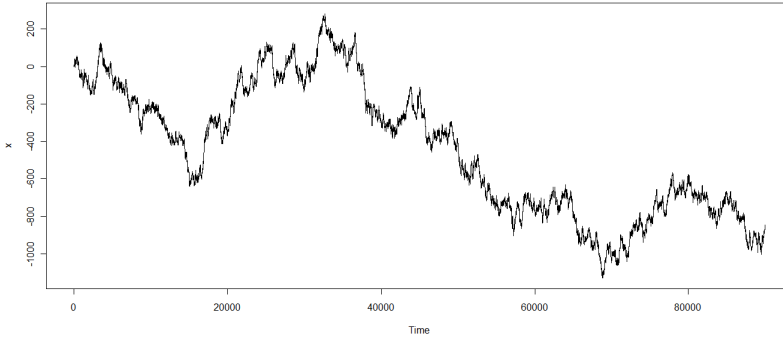
The evaluation of the public services is related to a specified set of different public policy alternatives: Marchitelli 2009 [486] and Drago et al. 2009 [230].

The aim of the analysis is to compare different policy alternatives that could be considered competitive. The data coming from evaluation procedures are difficult to measure and to model, so to face the problem of different measurement we need to use representations as intervals.

More precisely the problem can be considered an optimization one, where we want to choose the best alternative with respect to a specific metric and the budget constraint. The evaluation or the choice needs to take into account the local social development and the environment of a specific zone or territory. The evaluation can be ex-ante if the



(a) Interval and aggregated data



(b) The original time series

evaluation time is before the project development. Decisions are also constrained by a public budget and by actual regulations.

Ex post there is an evaluation based on a specific comparative analysis based on results we have obtained from a specific policy. The interest of the analysis is the ex-ante comparison of different projects or project options. All considered elements can contribute to the final outcome and the analysis. The analysis is related to the evaluation of

the Italian academic system and the analysis is divided into various steps²²:

1. Delivering a questionnaire on a non random sample of 40 academic professors of Italian universities participating in a workshop in Naples on the evaluation of academic institutions.
2. Each "judge" evaluates each scenario and each single component. The different evaluations are considered as a rank.
3. Obtaining some estimates from the Conjoint Analysis to have the utility coefficients.
4. From the utility coefficients to the symbolic data intervals by measuring the uncertainty.
5. The different judge coefficients are considered as interval data and interpreted through an Interval Principal Component Analysis.

The result shows that it is possible to use these data in a comparison of different policy alternatives.

1.5.2 Analysing the Financial Risk on the Italian Market using Interval Data

Drago and Irace in 2004 [231] 2005 [232] show the interval data to approach the financial analysis of the risk. In particular the idea is that of departing from a classical data set of financial data by considering their entire features to arrive at an analysis of the risk. Similar results for the French market using Symbolic Data are found in Goupil et al.

²²In particular we use the Factorial Conjoint Data Analysis: see Lauro 2004 [443]

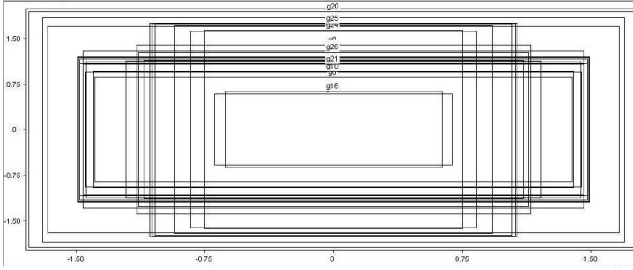


Figure 1.9: Judgement analysis and the interval symbolic data. Marchitelli 2009 [486] and Drago et al. 2009 [230]

2000 [317].

The data are related to the closing prices on the Italian stockmarket for a sample of listed companies. In particular the analysis is divided into three different parts. In the first part we perform a classical time series analysis to extract some patterns over time (in particular they can become the basis for the symbolic data analysis and the statistical arbitrage²³). Secondly we perform some multivariate methods (Principal Component Analysis, Statis, and Cluster Analysis), and thirdly we consider the intervals related to a single week, in which we consider the Interval principal component analysis and the clustering methods in interval data.

The analysis discriminates some interesting patterns regarding the market, and in particular stocks that tend to have higher returns with respect to their volatility etc.

At the same time intervals can show some opportunity of arbitrage over time (small intervals can show some profit opportunities)

²³In particular see Avellaneda Lee 2010 [50]

1.5. Applications on Real Data

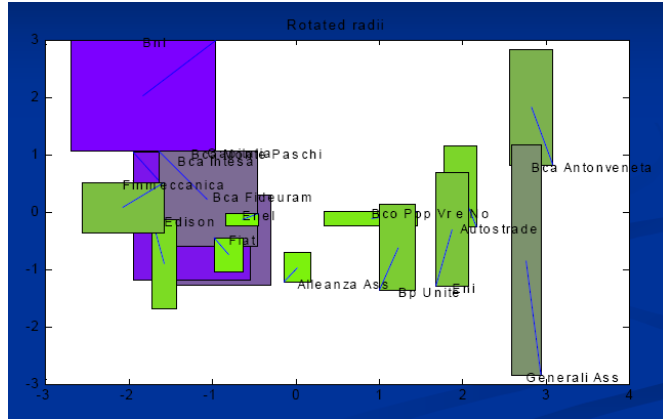


Figure 1.10: Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])

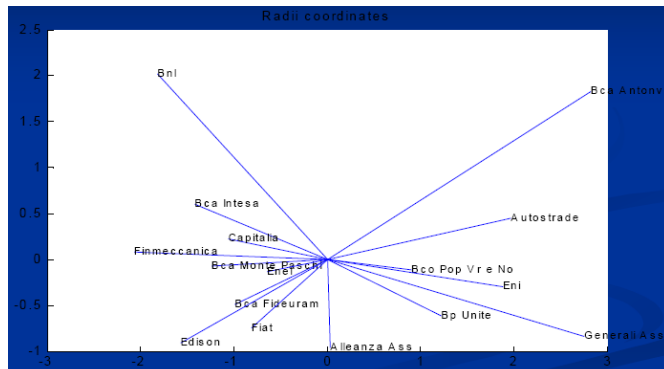


Figure 1.11: Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])

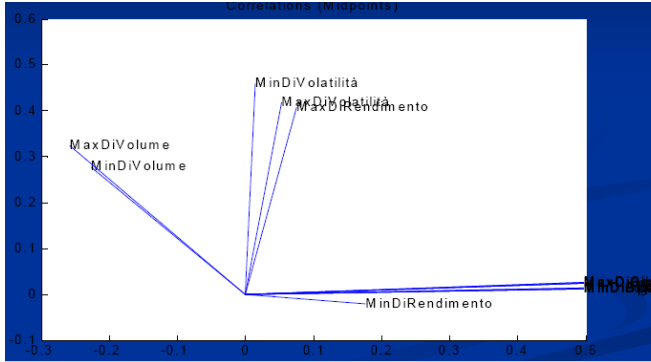


Figure 1.12: Interval Data Principal Component Analysis and Financial Data (Drago and Irace in 2004 [231])

Summary Results: The Analysis of Massive Data Sets
Massive and Huge data sets call for different methodologies which allows us to take into account not only the location of data but also its size and shape.
In Huge Data aggregation there is a loss of information.
Internal Representations (Intervals, Histograms, etc.) allow a more complete representation of the data.
Data can be genuinely considered to be Intervals or Histograms when they are characterized by complex patterns of variation.
In all the cases of the use of Internal Representations, we can represent the single data without any type of aggregation.
Symbolic Data are Internal Representations based on relevant assumptions.
These methods can be used in various fields, such as Policy Analysis and Financial Analysis.

Chapter 2

Complex Data in a Temporal Framework

The methods described in the thesis, as will be made clear later, can be applied in various data contexts, one of them being financial. So in this section we will present the characteristics of the high frequency financial data and at the same time the nature of these time series that call for the use of different methods¹.

We will motivate the interest in financial data² by showing the usefulness of the methods proposed in the thesis, by analysing some characteristics of the financial data as a whole and proposing them as

¹In particular in recent years there has been a growth of the use of different methods for financial data: see Tsay 2005 [667], Campbell Lo and McKinlay 1996 [118], Ruppert 2010 [599] Kovalerchuk and Vitayev 2000 [431], Tam Kiang Chi 1991 [650] and Mantegna Stanley 2000 [485] on the enormous literature of the quantitative methods in finance

²In particular, we refer to high frequency data that presents some unequal space characteristics which presenting important challenges to operators in the field, see: Dacorogna et al. 2001 [163], but also Zivot 2005 [722], Cont 2011 [154] and also Bauwens Hautsch 2006 [67]

an initial framework of our methods³ (see the discussion on internal and external modelling and their applications). It is important to note and underline that the methods can also be used in other frameworks or contexts⁴.

We analyse data at different frequencies⁵ because there is no best interval temporal in all circumstances and the generating process needs to be analysed (In this approach we follow directly Dacorogna 2001 [163]).

Here the term "complex" can be used in two different ways. The first one is in the sense that time series are typically original structured data coming from different sources (in this sense, see Diday 2010 [211], Diday Noirhomme 2008 [218], but also 2006 [208] and 2002 [207]) in which we can obtain some representations which summarize the temporal observation by retaining the meaning.

In the second meaning the time series could be characterized by some behaviors like irregular cycles, complex seasonal patterns, non stationarities, waves, peaks, nonlinearities outliers etc. (see De Livera Hyndman Snyder 2010 [190] Sewell 2008 [619] and Gao Cao Tung Hu 2007 [290]). In this sense some representations could be useful in detecting the underlying data structure.

In practice, in the complex time series we can assume its structure is hidden from the noise, so it is necessary to separate the "signal"

³It is important to note that a direct application of the method belongs to high frequency financial data for the intrinsic reasons of these data types, see for example Arroyo et al. 2011 [38] Drago Scepti 2010 [236], Drago Scepti 2010 [237] and Drago Lauro Scepti 2011 [235]

⁴Original complex data in temporal framework can be adapted to all disciplines, so the methods can be used in different frameworks (see for example Diday 1998 [205] on the definition of time series as complex data)

⁵An interesting methodology is MIDAS (Mixed Data Analysis) considering regressions by merging data from different sampling frequencies: Andreou Ghysels Kourtellis 2010 [20] and 2010 [21], Ghysels Sinko Valkanov 2007 [303] Ghysels (2005) [301]

from the "noise" to use the data in an improved way. We consider in Chapter 7 an approach to internal representations that takes into account model data, in which we will try to separate in the data the structural part from the noise⁶.

2.1 Homogeneous and Inhomogeneous Time Series

As seen in the first chapter there was an important evolution in financial markets due to the availability of new types of data. In particular in Financial Markets the advances in computer technology and data storage have made the high frequency data available for researchers (see in this sense Yan Zivot 2003 [710]). In particular, very important was the introduction of automated electronic systems of trading that permit the transformation into readily available of entire records containing the characteristics of all the trades and quotes executed in a regulated market (Galli 2003 [286]).

A typical example is the Trade and Quotes (TAQ) database⁷ that provides a collection of relevant information like prices and quotes for all the stocks listed on the New York Stock Exchange (NYSE), the American Stock Exchange, the Nasdaq National Market System (NASDAQ), and the SmallCap issues. An example of the final dataset is shown in Galli 2003 [286], see in particular figure.2.1.

So these data refer to time stamped transaction-by-transaction or tick-by-tick data, referred to as ultra-high-frequency data by Engle Russell (2004) [254]. There are known problems in time scales due to the occurrence of these new types of data (see Mantegna Stanley 2000

⁶A similar approach is chosen by Signoriello 2008 [630]

⁷See Kyle Obizhaeva Tuzun 2010 [440] for the characteristics of the database TAQ in trading game contexts

[485]). In specific financial contexts the number of observations in high frequency data sets can be overwhelming and data are often recorded with errors and there is a need for it to be cleaned and corrected prior to direct analysis.

Transaction by transaction data on trades and quotes are, by nature, irregularly spaced time series with random daily numbers of observations. High-frequency data typically exhibit periodic intra-day and intra-week patterns in market activity. In these cases the need for aggregation arises. In particular there is the need for techniques to summarize, visualize, cluster and forecast high frequency financial time series without information loss. At the same time, the need to study the financial market as a whole using multivariate tools and cointegrated time series where all data are high frequency data (there is no mix of different frequencies) is essential.

2.1.1 Equispaced Homogeneous Data

A scalar time series⁸ y_t can be represented along the notion of scalar stochastic process⁹. Following Peracchi (2001) [555]: on the space $\Omega \times \mathcal{T}$ can be defined a function Z as a scalar stochastic process, such that a random variable is defined on the probability space (Ω, A, P) for every $t \in \mathcal{T}$, $Z(-, t)$.¹⁰

Assuming a common probability space (Ω, A, P) a scalar stochastic process is a collection $\{Z(., t), t \in \mathcal{T}\}$ of random variables, a state

⁸In this thesis, we consider various types of time series, for example the scalar one (STS), but at the same time, interval time series (ITS), boxplot time series (BoTS), histogram time series (HTS) etc. See in this sense Arroyo, González-Rivera, Maté, San Roque (2011) [38] Arroyo González-Rivera Maté (2010) [41], Han Hong Wang (2009) [336]

⁹Ross 1996 [596] and Dobb 1953 [223]. Interval stochastic processes are defined in Han Hong Wang 2009 [336]

¹⁰Peracchi 2001 [555]

2.1. Homogeneous and Inhomogeneous Time Series

Time	Index	Type	Bid	Ask	Price	Volume
100405	36245	Q	10	$10\frac{1}{8}$		
100407	36247	T			10	500
100445	36285	T			10	700
100502	36302	T			$10\frac{1}{8}$	450
100506	36306	Q	$10\frac{1}{8}$	$10\frac{1}{4}$		
100507	36307	T			$10\frac{1}{8}$	100
100509	36309	T			$10\frac{1}{8}$	900
100513	36313	Q	$10\frac{1}{4}$	$10\frac{3}{8}$		
100610	36370	T			$10\frac{1}{4}$	2500
100611	36371	T			$10\frac{1}{4}$	250
100812	36492	Q	$10\frac{1}{2}$	$10\frac{5}{8}$		
100822	36502	T			$10\frac{1}{2}$	500
100824	36504	T			$10\frac{1}{2}$	200
100904	36544	T			$10\frac{5}{8}$	400
101547	36947	Q	$10\frac{3}{8}$	$10\frac{5}{8}$		
101548	36948	T			$10\frac{3}{8}$	1500
101550	36950	T			$10\frac{3}{8}$	700
101555	36947	Q	$10\frac{1}{4}$	$10\frac{3}{8}$		

Figure 2.1: High Frequency Data: Trades and quotes dataset (Galli 2003 [286])

space of the process can be considered the range of $Z(., t)$ ¹¹

The index space of a time series can be related to points in time $t_{i...j}$ that are equally spaced to a given time unit (Peracchi 2001 [555]).

A time series is a way to represent a stochastic process, it can be denoted as $y = (y_t : t \in T)$ could be studied in the time domain (for example autoregressive models as equations predicting y_t from y_{t-1} to y_{t-n}) for example, we can have in the case of an $AR(p)$ process:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (2.1)$$

Assuming:

$$\begin{aligned} E[\varepsilon_t] &= 0 \\ E[\varepsilon_t^2] &= \sigma^2 \\ Cov[\varepsilon_t \varepsilon_s] &= 0 \end{aligned} \quad (2.2)$$

If $\varepsilon \sim \mathcal{N}(\mu, \sigma^2)$ the process is Gaussian.

Data can be analysed by considering a specific frequency f (or temporal interval). There is no general need to aggregate different time series in considering different frequencies. A recent methodology that mixes data with different mixed frequencies is the MIDAS (Ghysels Kourtellis 2010 [20]) or mixed data sampling regression model¹².

Let us assume we have for two time series y_t and x_t and a frequency f :

$$y_t = \phi_0 + \phi_1 B(L^{1/f}; \theta) x_t^{(f)} + \varepsilon_t^{(f)} \quad (2.3)$$

¹¹A time series approach in space state is in Durbin Koopman 2001 [244] and Durbin 2004 [243]

¹²In another work, Ghysels Santa Clara Valkanov 2004 [302] consider the methodology also by taking into account high frequency data

2.1. Homogeneous and Inhomogeneous Time Series

In practice f denotes the frequency (daily, weekly, quarterly), $B(L^{1/f})$ is a lag distribution (for example, the Almon Lag). We start by considering the characteristics of high frequency data compared with equispaced homogeneous data.

A time series could be analysed, also, in the frequency domain (harmonic analysis, periodogram analysis and spectral analysis: see Battaglia 2007 [66]).

Box and Jenkins 1970 [99] show that the time-domain and the frequency domain show equivalent information on the time series¹³.

An important distinction needs to be made, in financial data in particular, by considering the spacing of the data points in time. Regularly spaced time series are usually defined as homogeneous whilst irregularly spaced are defined inhomogeneous (we use here the terminology in Dacorogna 2001 [163]). In the second case, we cannot apply standard methods that are designed for regularly spaced or homogeneous time series data¹⁴

In relation to the homogeneous time series we consider the finite time series x_t with length T and N observations. In the case of periodic sampling, the temporal distance between two realizations is always constant (Ng 2006 [535] and Zumbach Muller 2001 [726])

In fact, by considering t_i and t_j as single observations we have $t_i - t_j$ relating to the distance between two different observations, we have¹⁵:

$$t - t_j = \Delta t = \frac{1}{T} \quad \forall j \in \mathbb{N} \quad (2.4)$$

¹³An interesting review is proposed in Warner (1998) [693] focusing in particular on spectral analysis. A simple introduction to these methods is given by Brandes et al. (1968) [108]

¹⁴See Hamilton 1994 [333], Lutkepohl 2005 [469], Battaglia 2007 [66]

¹⁵Ng 2008 [536]

Ng 2006 [535] and in 2008 [536]¹⁶ shows that the time original homogeneous time series can be considered as a sum of trigonometric polynomials.

$$\begin{aligned}
 y_t &= \sum_{o=-N/2}^{N/2-1} a_o \cos(2\pi ot) + b_o \sin(2\pi ot) \\
 &= \sum_{o=-N/2}^{N/2-1} c_o \exp^{-i2\pi ot/N}
 \end{aligned} \tag{2.5}$$

In this case, the Fourier Coefficients can be computed by the Fast Fourier Transforms (also denoted as FFT).¹⁷.

$$c_o = \frac{1}{N} \sum_{t=1}^N x_t \exp^{-i2\pi ot/N} \tag{2.6}$$

High frequency transaction data arrives in irregular time intervals, the implementation of the common FFT cannot be done for the unevenly sampled data (Ng 2006 [535]).

In the case of the high frequency data we can consider the Point Process, we will consider it later (for a complete introduction to the Point Processes see Karr 1991 [418] and Daley Vere-Jones (1988) [165]).

2.1.2 Inhomogeneous High Frequency Data

High Frequency financial data are typically inhomogeneous. Many types of financial data can be obtained at high frequency, intraday

¹⁶See also Fricks 2007 [281]

¹⁷For a presentation of these methods see Percival Walden 2006 [556]

specifically, at a market tick by tick frequency. There are cases in which raw data are not suitable: in these cases market ticks arrive in random times. Engle (2000) [250], Zumbach and Muller (2001) [726] and more recently Zivot and Yang (2006) [723] deal directly with irregularly spaced data.

In that sense, following Zumbach and Muller (2001) [726] x_i is defined as an inhomogeneous time series. Here the market transactions over time can be characterized as (t_i, z_i) . In that sense t_i is the time and z_i is the scalar, representing, for example a price. So $z_i = z(t_i)$ and the point t_i are the i -th element of the series. More specifically an inhomogeneous time series is denoted as $(z_i)_i^N$.

In practice a time series x is defined by the arrival of ticks z_i at times t_i . Over time $t_{i+1} > t_i$. As we know, homogeneous time series are regularly spaced, inhomogeneous time series are irregularly sampled $t_{i+1} - t_i \neq \Delta t$ (Muller 1996 [520]).

Various methods can be considered for obtaining homogeneous time series from inhomogeneous (Dacorogna 2001 [163] Zumbach and Muller (2001) [726] and also Brownlees and Gallo 2006 [115]).

By starting with an inhomogeneous series with times t_i following Dacorogna 2001 [163], we have a transformation from inhomogeneous to homogeneous time series $z_i = z_{t_i}$. In practice only the last observation in the time is chosen. The sequence of the raw series can be related to the index i while in the case of the homogeneous we consider them at time t . By considering a specific period Δt we obtain a series: $t_0 + i\Delta t$ regularly spaced.

The ticks x_i indexed, each related to the market characteristics (see in figure 2.1 price and volume in particular) for each index i can be considered as the realizations of a marked point processes.

In particular, we can consider the different partitions of the TAQ dataset as a specific high frequency dataset, here we need to consider separately the Trades and the Quotes.

For trades tr considering $i = 1 \dots n$ as the number of the trades over

time we have $(tr_{1i}, tr_{2i}, tr_{3i})$, with tr_{1i} as the time of the trade tr_{2i} as the price and tr_{3i} as the volume.

For quotes qu , considering the number $j = 1...n$ at the same time we have $(qu_{1j}, qu_{2j}, qu_{3j})$ we have: qu_{1j} for the time qu_{2j} as the bid price qu_{3j} as the ask price (see Galli 2003) [286]. So Trade Duration data, in fig.2.2, can be finally defined:

$$D_i = tr_{1i} - tr_{1(i-1)} \quad (2.7)$$

Where we can define at the same time the Quote duration data (fig.2.3):

$$Q_j = tr_{1j} - tr_{1(j-1)} \quad (2.8)$$

These types of data can be analyzed by considering specific econometric methods, such as the ACD or Autoregressive Conditional Duration models¹⁸ (see Engle and Russell, 1998 see [253] and Engle 2000 [250]). In practice we model the durations between two distinct events¹⁹

$$duration_i = tr_i - tr_{i-1} \quad (2.9)$$

2.1.3 Irregularly Spaced Data as Point Processes

High Frequency Data present the particular characteristics of being inhomogeneous or irregularly spaced. In Statistics we define this type of process as a Point Process. In particular, a Point Process is a random element, whose values are in a defined set S . The outcome

¹⁸See also the interesting comparison between methods in Zhang Keasey and Cai [718]

¹⁹Forecasting the arrival of an event is clearly relevant in various other applicative contexts. See for example the forecasting methodology and its application in Shen and Huang 2008 [625]

2.1. Homogeneous and Inhomogeneous Time Series

Time	Index	Trade duration	Price	Volume
100407	36247		10	500
100445	36285	38	10	700
100502	36302	17	10	450
100507	36307	5	10	100
100509	36309	2	10	900
100610	36370	61	10	2500
100611	36371	1	10	250
100822	36502	131	10	500
100824	36504	2	10	200
100904	36544	40	10	400
101548	36948	404	10	1500
101550	36950	2	10	700

Figure 2.2: High Frequency Data: Trade durations (Galli 2003 [286])

Time	Index	Quote duration	Bid	Ask
100405	36245		10	10
100506	36306	61	10	10
100513	36313	7	10	10
100812	36492	179	10	10
101547	36947	455	10	10
101555	36955	8	10	10

Figure 2.3: High Frequency Data: Quote durations (Galli 2003 [286])

of the Point Process is an inhomogeneous time series, that could be converted to a homogeneous time series.

Following Hautsch 2007 [348] and Bauwens and Hautsch 2006 [67], given t as the time, a point process could be defined as W as a sequence of events w :

$$(t_i^w)_{i \in (1 \dots n_w)} \quad w = 1 \dots W \quad (2.10)$$

In this case, every event time of the pooled process becomes:

$$(t_i)_{i=1} \quad (2.11)$$

The inter event duration can also be:

$$d_i^w \equiv t_i^w - t_{i-1}^w \quad (2.12)$$

in a specific information set at $t : \mathcal{F}_t$. So a point process can be visualized as fig.2.3:

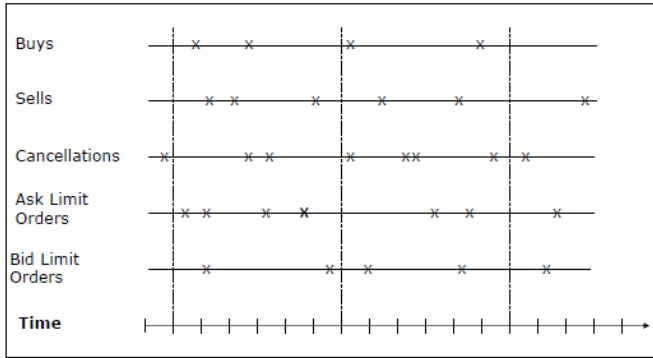


Figure 2.4: High Frequency Data: Point Processes (Hautsch 2007) [348])

The time that can be from the most recent event is:

$$x(t)^w \equiv t - t_{N^w(t)} \quad (2.13)$$

The information set can be defined t as F_t

Hazard function λ of a random variable X can be defined:

$$\lambda_X(x_i) \equiv \frac{f_x(x_i)}{(1 - F_X(x_i))} \quad (2.14)$$

At this point, it is necessary to observe the ways to transform inhomogeneous time series into homogeneous ones.

2.1.4 Inhomogeneous to Homogeneous Time Series Conversions

High Frequency Data related to a single stock can be affected by outliers, errors and anomalous values²⁰. The main technical reason for these problems is unknown. A preliminary step is clearly that of managing the outliers and anomalous observations which occur in the inhomogeneous time series (and we will see this later in financial high frequency data).

Following Dacorogna (2001) [163], Zumbach and Muller 2001 [726] and also Zivot and Yang 2006 [723] we can follow two ways to convert an inhomogeneous scalar time series into a homogeneous one. Denoting k' as a single tick and considering: the time $t_0 + i\Delta t$ we have:

$$k' = \max(k | t_k \leq t_0 + i\Delta t), t_{k'+1} \quad (2.15)$$

$$t_{k'} \leq t_0 + i\Delta t < t_{k'+1} \quad (2.16)$$

We want to interpolate between $t_{k'}$ and $t_{k'+1}$ using Linear interpolation:

²⁰Brownlees Gallo (2006) [115], Andersen 2000 [16]

$$z(t_0 + i\Delta t) = z_{k'} + \frac{t_0 + i\Delta t - t_{k'}}{t_{k'+1} - t_{k'}}(z'_{k+1} - z'_k) \quad (2.17)$$

and the previous (the most recent) tick, interpolation:

$$z(t_0 + i\Delta t) = z_{k'} \quad (2.18)$$

where it is also possible to consider the initial (the oldest) tick interpolation:

$$z(t_0) = z_{t^0} \quad (2.19)$$

In practice, various other interpolation functions can be used but the previous tick interpolation is the most used.

In this way we can obtain a homogeneous time series. The result is a loss of information related to the intra-day dynamics. This loss can be higher if the temporal interval chosen is higher (see Engle 1996 [250]). In particular the analysis becomes difficult for the complex form of strong intraday seasonalities (for example Fantazzini and Rossi 2005 [261]. At the same time, interpolation, in the presence of non synchronous trading, can introduce spurious correlations (see Engle Russell 2004 [254]).

Another problem of the aggregated data is cited by Dacorogna 2001 [163] and is related to the fact that using aggregated data can be dangerous because of the presence of changes of the indexes used in the data and the impact of these changes is not predictable. At the same time, using the high frequency data brings new problems because the seasonality of these data can hide other data structures.

2.2 Ultra High Frequency Data Characteristics

In recent years there has been an increase of the availability of higher frequency measurements of the economy. This fact has been very positive because it has made for the improvement of the likelihood of the statistical analysis²¹. It is important to note that the limit of this process can be reached when all the transactions have been recorded and this could happen in different markets. In this sense, each market can generate transactions, so data could be collected in the same time tick by tick (or observation by observation) by considering different products in the market locations such as supermarkets, internet, or financial markets, see Engle 2000 [250].

So, high frequency data, as inhomogeneous time series, are increasingly important in financial markets. They allow us to study the adjustment processes of prices in the financial markets, the intraday dynamics and the market microstructure (for the mechanisms in which the new information impact on price see for example Cont 2011 [154] and Engle Russell 2004 [254]²². High-frequency data are also useful for analysing, at a lower frequency, the volatility of asset returns and for portfolio allocation²³. At the same time, just as the high frequency data has fuelled advanced forms of algorithmic trading so has the high frequency trading, based on these data²⁴.

The first set of characteristics we will analyse in depth is: irregular temporal spacing (as a characterising difference with other types of

²¹See for example Modugno 2011 [511]

²²In particular the relationships between "order flow, liquidity and price dynamics" in Cont (2011) [154]. See also Bouchaud Mézard Potters 2002 [104] Farmer et al. 2004 [264]

²³Corsi, Dacorogna, Müller, Zumbach (2001) [159], Barucci Renó (2002) [61] and Hautsch Kyj Malec (2011) [349]

²⁴Ahmed M. 2009 [10] and Patterson Rogow 2009[552]

financial data at a daily or weekly frequency), diurnal patterns, price discreteness, and very long dependence (Engle Russell 2004 [254], Dacorogna 2001 [163]).

Another important reason is to understand how the financial microstructure can affect price dynamics²⁵.

It is important to note that the institutional regulations that collect the information on markets (also defined as the market microstructure) can change data characteristics. Changes in market regulations and technological advances modify the data characteristics and the different data structures. For an example of this variability over the time and the space of the data characteristics see Brownlees and Gallo 2006 [115].

2.2.1 Overwhelming number of observations

Size refers to the number of ticks in a specific trading day. The high-frequency databases storing tick-by-tick data are very complex to analyze, for example Brownlees, Gallo 2006 [115]. The first characteristic for the high frequency financial datasets is that they are of an overwhelming size.

The transmission by data vendors can vary from market to market, for example Reuters for a foreign exchange spot rate distributes more than 75,000 prices per day (see Dacorogna 2001 [163]). In other cases, the datasets size can be over 10 million foreign exchange price quotes (Dacorogna Muller Pictet De Vries 2001 [164]).

Clearly the different sizes of the data sets depend on the different markets²⁶. Mykland and Zhang (2009) [525] cite the fact that on a

²⁵Cont (2011) [154] Biais, Glosten, Spatt (2005) [79] and Rosu 2009 [597]

²⁶For an example of the specific characteristics of high frequency data for exchange rates see Dacorogna et al. 1993 [162] and 1990 [161]. Dacorogna Muller Pictet De Vries 2001 [164] study the exchange rates focusing on the outliers and the heavy tail features of foreign exchange returns. The authors found at the

single day Merck had 6,302 transactions and Microsoft had 80,982. In general, we can conclude that the size is very relevant.

At the same time, the number of ticks, or observations, produced can vary greatly considering the different markets and different financial instruments. A “highly traded stock may have tens of thousands of price events per day, quickly resulting in a storage requirement to store Gigabytes of data per day and Terabytes of data per year for any reasonable sized instrument universe” (Xenomorph 2007 [706]).

Engle and Russell confirm that these data “are characterized by ten thousands of transactions or posted quotes in a single day time stamped to the nearest second” [254]²⁷.

It is important to note that if the data volume is growing exponentially so at the same time the storage of the intraday or high frequency data is becoming difficult²⁸.

This type of characteristic neglects for example to visualize correctly the data and it is necessary to consider distinct windows for observing all data²⁹ (see Drago and Scepti [237]).

2.2.2 Gaps and erroneous observations in data

High Frequency datasets contain gaps and wrong observations, and some unordered sequences³⁰. At the same time, Brownlees and Gallo 2006 [115] report that these data can be characterized by some anomalous and occasional behavior determined by specific market conditions (opening, closing, trading halts, for example). In particular, the 2-

same time that high frequency data improve the efficiency of the tail risk cum loss estimates

²⁷see also Yang Zivot 2003 [711] and Falkenberry 2002 [260]

²⁸See in the specialized journal Automatic Trader the interview of Brian Sentance (2007) [732]

²⁹Zivot 2005 [722]

³⁰Yang Zivot 2003 [711]

3%³¹ represents erroneous ticks that need to be imputed when they can determine bad trades. There is in that sense a call for real time filtering and data cleaning algorithms. At the same time, there is a problem of overscrubbing data³².

It is necessary to take into account this phenomenon and to clean the data before the analysis. In particular, classical econometric and statistical techniques do not permit the solving of these problems. Some techniques to handle these types of problems, in particular outliers, are proposed by various authors: Dunis et al. 1998 [241], Brownlees and Gallo 2006 [115] and Mineo and Romito 2008 [508]

There are two steps in the method, the first step is the transformation of inhomogeneous time series into a homogeneous one. The second step is data filtering or cleaning: here it is necessary to detect the outliers and to manage them in a suitable manner.

Various proposals in this sense are to be found in literature: Dacorogna 2001 [163] for the Olsen & Associates algorithm and, Zhou 1996 [721]. Following the proposal of Brownlees and Gallo 2006 [115] let $(z_i)_i^N$ be a high frequency time series (an ordered tick by tick series), in the example related to a price we have:

$$(|z_i - z_{\text{trimmed mean}_i(d)}| \leq 3s_i(d) + \varsigma) = \begin{cases} \text{TRUE } i \text{ is kept} \\ \text{FALSE } i \text{ is removed} \end{cases} \quad (2.20)$$

In that case, $z_{\text{trimmed mean}_i(d)}$ and $s_i(d)$ are the α trimmed mean and the standard deviation of a neighborhood of d observations around i and ς is a granularity parameter³³ If the expression is true the i ob-

³¹Falkenberry 2002 [260]

³²Falkenberry 2002 [260]: an introduction to the problems of data cleaning and so to underscrubbing and overscrubbing data, removing the volatility structure (Dacorogna 2001 [163])

³³Mineo and Romito 2008 [508] explains: "The granularity parameter is con-

servation is kept, if the expression is false the i observation is declared an outlier and it is removed.

Mineo and Romito 2008 [508] make a proposal in this sense:

$$(|z_i - z_{mean_i}(d)| \leq 3s_{-i}(d) + \varsigma) = \begin{cases} \text{TRUE } i \text{ is kept} \\ \text{FALSE } i \text{ is removed} \end{cases} \quad (2.21)$$

In that case $z_{mean_i}(d)$ and $s_{-i}(d)$ are the mean and the standard deviation of a neighborhood of d observations around i without the i -th observation and ς is the granularity parameter. A third approach considered in Dunis (2008) is related to the median of the last three ticks³⁴

In this sense, the outliers are detected and removed from the data. For a specific guide to the data analysis preparation and variable creation see Yang Zivot 2003 [711]. See in figure 2.5 and figure 2.6 some examples of erroneous observations in high frequency data.

2.2.3 Price discreteness

There are important features in high frequency data that determine some spurious auto-correlations in data, for example, the price discreteness³⁵. The price discreteness in the high frequency data is related to the high kurtosis manifested from data³⁶.

Price discreteness can be considered to be the truncation of prices

sidered because the ultra high-frequency series often contains sequences of equal prices which would lead to a zero variance; thus, it is useful to introduce a lower positive bound on price variation in order to have always admissible solutions.

³⁴Mineo and Romito 2008 [508] report that the Dunis method performs with the worst performance in respect to other methods used

³⁵Matei 2011 [495]

³⁶Dacorogna 2001 [163] and Engle Russell 2004 [254]



Figure 2.5: Erroneous observations in High Frequency Data: one spike (Browne 2011 [746])

or exchange rates in a small number of digits, with respect to an infinite number of digits.³⁷ Engle, and Russell, wrote: "All economic data is discrete. When viewed over long time horizons the variance of the process is usually quite large relative to the magnitude of the minimum movement. For transaction by transaction data, however, this is not the case and for many data sets the transaction price changes take only a handful of values called ticks".

At the same time the market regulations and the institutional roles can have a role in restricting the prices to fall on a specific set of values. Engle Russell 2004 [254] states:" price changes must fall on multiples of the smallest allowable price change called a tick. In a market for an actively traded stock it is generally not common for the price to move

³⁷McGroarty, Gwilym, Thomas (2006) [497]

2.2. Ultra High Frequency Data Characteristics



Figure 2.6: Erroneous observations in High Frequency Data: bid ask gapping (Browne 2011 [746])

a large number of ticks from one transaction to another.”

There are differences between markets due to internal characteristics: ”In open outcry markets the small price changes are indirectly imposed by discouraging the specialist from making radical price changes from one transaction to the next and for other markets, such as the Taiwan stock exchange these price restrictions are directly imposed in the form of price change limits from one³⁸”.

2.2.4 Seasonality and Diurnal patterns

Intraday financial data usually contain relevant diurnal or periodic patterns (see Sewell 2008 [619], Engle Russell [254] 2004 and Dacorogna 2001 [163]). As Engle Granger [254] wrote, it is possible

³⁸Engle Granger 2004[251]

to detect a U-shaped pattern over the day, for most stock markets, volatility, the frequency of trades, volume, and spreads show a U-shaped pattern.

These patterns need to be considered before beginning any statistical inference on the data³⁹. In that sense, it is important to explore the data to detect the pattern of seasonality in data.

In any case, various different patterns can be detected in the markets⁴⁰ and they can be determined by complex relations, for example with market microstructure or the public arrival of information.

These patterns are clearly different from the patterns we are able to extract in financial data at a lower frequency, as for example, the well known "January effects"⁴¹. In these cases we can define them as effects due to market anomalies that challenge the idea of perfect efficiency in the financial markets. In any case, the idea of data structures and different patterns in different types of data is confirmed by considering different data frequencies⁴².

2.2.5 Long dependence over time

Following the approach of the data characteristics of Engle and Russell 2006 [254] related to the market structures, the authors refer to various phenomena: the existence of a long dependence over time is one of the most important.

The dependence can be considered the result of price discreteness and the spread between the price paid by buyer and seller initiated trades. Long dependence can be referred to as bid-ask bounce and the large first order negative autocorrelation.

³⁹Melvin Yin (2000) [503] Andersen and Bollerslev (1994) [17] and Dacorogna, et.al (1993) [162]

⁴⁰Kunst 2007 [439]

⁴¹Keim 1983 [420]

⁴²Dacorogna 2001 [163]

Traders breaking large orders up into a sequence of smaller orders in the hope of transacting at a better price overall can lead to a dependence in price changes. These buys and sells, so sequentially ordered, can determine a sequence of transactions that change prices in the same direction.

Hence, on long term horizons we sometimes find positive autocorrelations. This result is confirmed in many studies of intra day data⁴³

2.2.6 Distributional characteristics and Extreme Risks

See Dacorogna 2001 [163]. Distributional characteristics change with respect to the frequency, the more frequent the data the more the data share high frequency characteristics, the more the distributions are fat tailed. The general result for the distribution is that the data are fat tailed [163] and also characterized by strong skewness [261]. In general, by considering the choice of little intervals of time for the conversion to inhomogeneous time series of homogeneous ones (for example 15 minutes) the distribution chosen is leptokurtic (see also the nonparametric approach in Coroneo Veredas (2006) [157].

The presence of high frequency data allows for the good analysis of fat tails. From the applicative point of view it is interesting to analyse the tails to understand the extreme movements in the financial markets. The Extreme Risk analysis is a growing area in the field of Financial Econometrics (see Cont 2001 [152]).

2.2.7 Scaling Laws

At the same time scaling laws are present in financial data at different frequencies (see in particular Dacorogna 2001 [163] and Sewell 2008

⁴³Sun Rachev Fabozzi 2007 [646]

[619]). Following Sewell 2011 [621] who reviews the characteristics of the financial time series "Scaling laws describe the absolute size of returns as a function of the time interval at which they are measured. Markets exhibit non-trivial scaling properties".

At the same, Sewell in two works in 2011 [761] and [621], reviews various different contexts in which it is possible to find scaling laws in finance irrespective of underlying data or frequency.

So, there now exists a vast empirical verification from that of the initial work of Muller et al. 1990 [520] which found the existence of scaling laws in financial data like the FX rates (see Dacorogna 2001 [163]). The evidence was confirmed also by considering other markets using different financial instruments.

2.2.8 Volume, Order Books and Market Microstructure

The dynamics of the volumes, order books, and market microstructure, seem to be relevant in understanding how markets work in reality. At the same time, the market functioning, market institutions and market processes impact on trading costs, prices, volume and trading behaviour (see Sewell 2008 [619] but also Tsay 2005 [667]).

In particular, Tsay 2005 [667] refers to High-frequency data as a key to understanding some characteristics useful in analysing and understanding some phenomena like:

1. Nonsynchronous trading
2. Bid-ask spread
3. Duration models
4. Price movements that are in multiples of tick size

5. Bivariate models for price changes
6. Time durations between transactions associated with price changes

In general, the advantage of using high frequency data gives the possibility of investigating the concrete elements that relate to the determinants of the adjustment price.

High frequency data can be very relevant in the analysis of the different trading processes in various markets related to the market microstructure. In particular (see Tsay 2005 [667]) they can be used to compare the efficiency of the trading systems in price discovery. The analysis of market efficiency can be carried out by considering high frequency data⁴⁴.

The use of the high frequency data is important, here, for the results that are new in respect to the low frequencies.

2.2.9 Volatility Clustering

Similarly to financial data at lower frequencies we can observe the volatility clustering phenomenon at higher frequencies.

In particular, we can take into account the seasonal heteroskedasticity by considering the daily and the weekly clusters of volatility (see Dacorogna 2001 [163]).

In any case, this observation allows us to investigate if we can observe the same data characteristics in different data frequencies. Now, considering the data at a different frequency (a lower one) we investigate its specific characteristics⁴⁵.

⁴⁴Lillo 2010 [456] Tiozzo 2011 [659]

⁴⁵Sewell 2011 [621]

2.3 Financial Data Stylized Facts

Whereas High Frequency Data are a new data type, that promise to enlighten various phenomena on the financial markets, the classical way to consider financial data is a lower frequency (daily and weekly data).

On the general characteristics of the financial data at a lower frequency (days) there is a growing literature based on the so called stylized facts (see Sewell 2008 [619], Cont 2001 [152], Tsay 2005 [667] also in the new field of the Econophysics Mantegna and Stanley 2000 [485]).

Here we focus on the statistical characteristics of the financial time series. There are two types of series that are generally used: price series p_t^f at a given frequency f and return series r_t^f . p_t^f can be considered as the price of a financial item, for example an asset, and the $\ln(p_t)$ as its logarithm transformation, useful for many purposes⁴⁶. The r_t^f , with respect to p_t^f shows useful statistical properties, so it is widely used, in particular we have (see Tsay 2005 [667]) the simple gross return:

$$1 + r_t^f = \frac{p_t^f}{p_{t-1}^f} \quad (2.22)$$

or also

$$p_t^f = p_{t-1}^f (1 + r_t^f). \quad (2.23)$$

The simple net return is:

$$r_t^f = \frac{p_t^f}{p_{t-1}^f} - 1 = \frac{p_t^f - p_{t-1}^f}{p_{t-1}^f} = r(t, T) \quad (2.24)$$

⁴⁶See Lutkepohl Xu 2009 [471]

Following Cont (2001) [152], and generalizing for different temporal scales in a defined time scale Δt , that could be considered as a second or a month in a homogeneous time series, the log return at scale Δt will be:

$$r_{t,\Delta t} = x_{t+\Delta t} - x_t \quad (2.25)$$

By considering the returns of a portfolio q of assets l with weights α_i on the different assets, given the simple return of an asset $r_{l,t}^f$, we can define the simple return of the portfolio q , at time t as:

$$r_{q,t}^f = \sum_{i=1}^N \alpha_i r_{i,t}^f \quad (2.26)$$

See Tsay (2005) [667].

2.3.1 Random Walk Models and Martingale Hypothesis

A relevant hypothesis used in an important class of financial data (the prices) is the Random Walk and the Martingale Hypothesis. By following Tsay (2005) [667] Samuelson (1965) [605] and Mantegna Stanley (2000) [485] and Campbell Lo MacKinlay Lo (1996) [118] in general, the hypothesis of market efficiency is related to the p_{t+1}^f the price as the past values $p_0^f, p_1^f, \dots, p_t^f$ through the conditions:

$$E(p_{t+1}^f \mid p_t^f, p_{t-1}^f, \dots) = p_t^f \quad (2.27)$$

$$E(p_{t-1}^f - p_t^f \mid p_t^f, p_{t-1}^f, \dots) = 0 \quad (2.28)$$

Stochastic processes like these are defined martingales. There is no possibility for making profits. The concept is linked with the random walk price model, given by:

$$p_{t+1}^f = p_t^f + \eta_t \quad (2.29)$$

It is not possible to forecast the difference $p_t - p_{t-1}$ so the best prediction for p_{t+s} is p_t (see Nakamura Small 2007 for the tests associated with the Random Walk Hypothesis [528]). A Random Walk with drift became:

$$p_t^f = p_{t-1}^f + \beta + \eta_t \quad (2.30)$$

where an upward trend appears considering $\beta > 0$.

$$p_t^f = p_0^f + \beta t + \sum_{t=1}^n \eta_t \quad (2.31)$$

$p_0 + \beta t$ is the deterministic trend, where $\sum_{t=1}^n \eta_t$ is the stochastic trend. For the properties see Mantegna and Stanley 2000 [485]. In the weak form of the market efficiency there is no simple way to use past information in order to gain a profit. Financial data show a complex structure because they convey a large quantity of mechanisms that overlap in the influence of the series, the objective is to separate the information that is possible to predict, that could be evidence that markets are not completely efficient (see Mantegna and Stanley 2000 [485]).

One objective for the representation considered during the thesis is to discover some useful structures, both for the high frequency data and the classical financial data. In this sense, these market inefficiencies could be exploited (until the markets recover efficiency). For a list of market inefficiencies see Sewell 2008 [619], Lo and MacKinlay 1999 [466] and for the contrary opinion see Malkiel 1973 [481]. In recent years many financial models have used data sources as high frequency data (see Dacorogna 2001 [163]).

2.3.2 Distributional Properties of Returns: Fat Tails

Cont (2001) [152], defined the joint distribution of the returns as $r(t, T)$. The unconditional distribution of returns can be defined as well as:

$$F_T(u) = P(r(t, T) \leq u) \quad (2.32)$$

The kurtosis can indicate the deviation from the normal distribution:

$$\kappa = \frac{r(t, T) - (r(t, T))^4}{\sigma(T)^4} - 3 \quad (2.33)$$

where $\sigma(T)^4$ is the variance of the log returns $r(t, T) = x(t + T) - x(t)$. The kurtosis is defined such that $\kappa = 0$ for a Gaussian distribution, a positive value of κ indicating the fat tail, or the slow asymptotic decay of the PDF. In this way it is possible to take into account the risk. Cont (2001) [152] defines the Value-at-Risk (VaR) as a "high quantile of the loss distribution of a portfolio over a certain time horizon":

$$P(W_0(r(t, \Delta) - 1) \leq VaR(p, t, \Delta)) = p \quad (2.34)$$

Where W_0 is the present market value of the portfolio, $r(t, \Delta t)$

its (random) return between t and $t + \Delta t$. Δt is typically taken to be one day or ten days and $p = 1\%$ or 5%

2.3.3 Heterogeneity and Structural Changes

We follow the characterization of the financial time series proposed by Sewell (2011) [621]. In that sense, financial returns $r(t, T)$ are typi-

cally non stationary⁴⁷. In fact the standard deviation of the returns tend to be not stationary over time. Following Mantegna and Stanley 2000 [485] the same volatility on markets is time dependent.

Financial data shows frequent structural changes. This result is coherent with the idea that the structural changes need to be approached in the analysis using different time windows for forecasting or modelling using the data (see Pesaran and Timmermann 2004 [558]). Markets are continuously changing so parameter drifts in the models can occur frequently.

In general the ARCH (Engle 1982 [249]) and GARCH (Bollerslev [102] 1986) models for returns $r(t, T)$, frequently used in finance are non stationary in variance but not in mean ⁴⁸. At the same time a structural change occurs in a given defined statistical model, for example:

$$p_t^f = \beta + \beta t^f + \epsilon_t, t = 1 \dots T \quad (2.35)$$

with $\epsilon_t \sim iidN(0, \sigma^2)$ No structural change parameters $\alpha = 1.2$ and $\beta = 1$ Structural change $\beta = 2$ for $t > 150$

Various tests in the literature exist to detect the existence of structural change in financial data. The best known are the Chow Forecast test (Chow 1960 [140]) and also the CUSUM and CUSUMSQ Tests (Brown Durbin Evans 1975 [114]).

2.3.4 Non-Linearity

Various types of nonlinearity can be detected in financial stock returns $r(t, T)$. Sewell in his work [621] reviews the results in literature related to the phenomenon. Financial markets show not only frequent

⁴⁷Sewell (2011) [621], for a different approach Starica and Granger 2005 [640] De Lima 1998 [181]

⁴⁸Sewell 2008 [619]

structural changes but at the same form some predictable forms of nonlinearities (see Lim Brooks Hinich 2008 [457]).

Typically the time series models can be non linear in mean and or in variance (as the ARCH or GARCH models)⁴⁹.

2.3.5 Scaling

Scaling can be defined as a relation between time intervals t and the average volatility measured at a power η of the absolute returns observed (Sewell 2008 [619]). From the first work of Mandelbrot in 1963 [482] that found scaling, in cotton prices p_t , various other works have found scaling in financial data⁵⁰. It is possible to confirm the same observations seen in paragraph 2.2.7 on scaling laws related to the high frequency financial data. In that sense, it is possible to conclude that a predefinite data frequency for this financial time series characteristic does not exist.

2.3.6 Dependence and Long Memory

Following Cont 2001 [152] and Sewell 2008 [619].

If time lags are denoted as τ : the correlation between the different lags s becomes:

$$\text{corr}[r(s + \tau, \Delta t), r(s, \Delta t)] \quad (2.36)$$

By assuming the hypothesis of "market efficiency" (the "efficiency market hypothesis" depicted in 2.3.1) then it can be hypothesized that there is no autocorrelation of the returns $r(t, T)$.

In any case the "market efficiency" hypothesis seems to be too strong in some markets and some "market inefficiencies" can appear

⁴⁹Engle 1982 [249] and Bollerslev 1986 [102]

⁵⁰Sewell 2011 [621]

(and their use is a matter of empirical investigation⁵¹). Some examples in literature of autocorrelation of returns can be found in Sewell 2011 [619].

A different phenomenon is the Long Memory dependence of daily stock return series. In this case, the evidence (for returns, volatility, volume etc.) is mixed by using the R/S Statistic, also known as "rescaled range" or "range over standard deviation" from Hurst 1951 [374]. Clearly a Hurst Exponent value (or R/S statistic) related to the:

$$H(0.5 < H < 1) \quad (2.37)$$

can suggest an inefficiency in the considered market⁵² (where we can identify a long memory process).

Following Cont 2001 [152] given a time scale Δ the log return of the scale Δ is given by $rt = X_{t+\Delta} - X_t = \ln(\frac{S_{t+\Delta}}{S_t})$

A stationary process Y_t (with finite variance) is said to have long range dependence if its autocorrelation function $C(\tau) = \text{corr}(Y_t, Y_{t+\tau})$ decays as a power of the lag τ ⁵³ : $C(\tau) = \text{corr}(Y_t, Y_{t+\tau})_{\tau \rightarrow \infty} \sim \frac{L(\tau)}{\tau^{1-2d}} \quad 0 < d < \frac{1}{2}$

2.3.7 Volatility Clustering

The phenomenon of volatility clustering, is related to the fact that in financial time series: "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes" (Mandelbrot 1963 [482]).

The phenomenon are usually considered by taking into account the ARCH and GARCH models, in which volatility is related to the last

⁵¹Lo Mackinlay 1999 [466]

⁵²Peters 1996 [560]

⁵³Cont 2001 [152]

period volatility (period of calm on markets followed by periods of turbulence).

2.3.8 Chaos

Chaos can be defined (Sewell 2008 [619]) when it is possible to detect in data an unpredictable long term behavior that could be generated by some sensitive initial conditions in a deterministic dynamical system. In practice the behavior of a chaotic time series is usually non distinguishable from another stochastic time series (Barnett Salmon Kirman (eds.) 1996 [60]).

Sewell in 2011 [621], reviews various works that test the existence of chaos in data by concluding that there is little empirical evidence in financial markets of low-dimensional chaos.

2.3.9 Cross Correlations Between Assets

To understand the correlation between assets see Cont 2001 [152] and Tsay 2005 [667].

In particular, it is very important to consider the problems when we deal with many assets. In various financial applications and problems (for example portfolio asset allocation and risk management) it is very relevant to work not only with single assets.

In this sense, the joint distribution of the returns of the assets need to be known in order to conduct a statistical analysis of these data: Cont 2001 [152]. An important outcome in these types of analyses is the understanding of the contagion mechanisms of different stocks (or markets) in the crises. These types of correlations vary over time and space.

Summary Results: Complex Data in a Temporal Framework
--

Financial Data show relevant characteristics, which could be considered for Internal Representations.

High Frequency Financial Data seems to be problematic, for example in data visualization.

There is the need for appropriate techniques to extract complex patterns from the data.

Chapter 3

Foundations of Intervals Data Representations

We have seen in Chapters 1 and 2 that sometimes Internal Representations can be useful in retaining information from overwhelming datasets. That is, it is useful to consider the entire data structure or the entire distribution of the data.¹ In that sense, the existent literature in Data Analysis using Internal Representation of big data² uses mainly two classes of approaches: a first one using Interval data (analysed in this chapter) and a second one using Histogram and or Boxplot

¹See in Williamson 1989 [701] "If one looks at the development of the measurement process during the past century one soon observes that with increasing frequency the raw data are probability distribution functions or frequency functions rather than real numbers This is so in the physical sciences and in the biological and social sciences it is the rule rather than the exception. One may thus convincingly argue that distribution functions are the numbers of the future and that one should therefore study these new numbers and their arithmetic (Berthold Schweizer)".

²For example the literature in Symbolic Data Analysis, see in that sense Billard 2010 [82] and Diday Noirhomme 2008 [218]

data (analysed in Chapter 4)³. In the recent past there has been an evolution using other instruments, based on different approaches. In Chapter 5 we propose a new method of analysis of big data based on Kernel Density Estimation. At the same time, it is relevant to note that this chapter could be a mathematical foundation of the densities as internal representation. The densities or the beanplot we will look at in Chapter 5 use upper and lower bounds to understand relevant phenomena. So this chapter starts with the mathematical foundations of the internal representations (IR).

We stress that these representations can be genuine⁴ or can come from previous statistical processes that transform the initial data matrix into representation matrices⁵. At times we can obtain our original data as intervals, histograms etc. The classical data type is used when taking into account the imprecision⁶, the interval is useful to consider a range of different values in temporal aggregations in the interval. In this chapter the foundations of the interval analysis and its algebra, and the interval random variables, are presented. Starting from this it is possible to develop the interval stochastic processes and interval time series (ITS). In the next chapter we symmetrically develop the theory for the boxplots and the histogram data (another representation). In any event it is important to start from probabilistic arithmetic (Williamson 1989 [701]) that can be considered relevant in

³A general foundation of these approaches can be found in the imprecision and the vagueness, some data can be measured considering uncertainty (or risk) and so can be considered intervals, boxplots, histograms etc. to measure such uncertainty. See for an introduction to this approach: Palumbo (2011) [544]

⁴In that case there can be a specific interpretation of these data that could be found in Nature. An example is temperature with its range of minimum and maximum-considered a genuine interval data

⁵For example interval data can come from database queries see for example Diday 2002 [206]

⁶See for example in Gioia 2008 [309] the classification of error that could be found in data and in their solutions

order to understand the interval data (or the upper and lower bound of the density data).

3.1 Internal Representation Data and Algebra: intervals

3.1.1 Probabilistic Arithmetic

The idea in Williamson 1989 [701] is to calculate the distribution of arithmetic functions of random variables. In some cases the results are obtained in terms of dependency bounds (for example, in the case of the lower and upper bounds). In practice, the probabilistic arithmetic is a generalization of interval arithmetic that considers only the supports of the distributions of the variables⁷. The problem addressed⁸ is given by considering $Z = L(X, Y)$ with X and Y as independent random variables, where its distribution function is $F_{X,Y}$. The distribution F_Z function of Z could be written:

$$F_Z(z) = \int_{L(z)} d(F_X(u)F_Y(v)), \quad (3.1)$$

where: $L(z) = ((u, v) \mid u, v \in \mathfrak{R}, L(u, v) < z)$. The author states that to compute the expression is a necessary but not a sufficient condition for a probabilistic arithmetic. The dependency error can be defined as the error in computing the distribution of some function of V and W ($U = V/W$) assuming V and W are independent. For example X , Y and Z are assumed independent random variables, but $V = X + Y$ or $W = X \times Z$ are not necessarily independent.

⁷Williamson 1989 [701]

⁸We refer on Williamson 1989 [701]

Williamson 1989 [701] states that is possible to obtain partial solutions to these problems in terms of dependency bounds or the lower and upper bounds on the distributions of functions of random variables when only the marginal distributions are known. So we have: ldb and udb as lower and upper dependency bounds, where \square is a binary operation, it is possible to write the bounds:

$$ldb(F_X, F_Y, \square)(z) \leq F_Z(z) \leq udb(F_x, F_y, \square)(z) \quad (3.2)$$

At the same time the general approach has been studied by various authors. Springer 1979 [637]⁹ considers the idea of an algebra of random variables to calculate the convolutions. Various authors have also studied the numerical methods for calculating distributions of functions of random variables¹⁰. In the case of interval data we are more interested in the bounds of a distribution within an interval.

3.1.2 Interval Data and Algebra

In particular, Interval Data are the simplest and used means to represent both the intra-period and the measurement error or imprecision.

Various contributions¹¹ have appeared from the original work of Sunaga 1958 [647], and Moore 1962 [512] and 1966 [513]. At the same time, interval data and interval arithmetic found a relevant application obtaining reliable simulation mechanisms, see in this sense Batarseh

⁹The idea was applied in Downs Cook Rogers 1984 [226]

¹⁰Williamson 1989 [701] in particular in the Chapter 2

¹¹Gioia 2008 [309] reviews the different approach in Interval Algebra from the first contributors: Burkill 1924 [117], Young 1931 [712], Warmus 1956 [691] then Sunaga 1958 [647] and Warmus 1961 [692]. At the same time, modern approaches by Moore 1966 [513], Alefeld-Herzerberger 1983 [12], Kerarfott-Kreinovich 1996 [423], Neumaier 1990 [533], and Alefeld-Mayer 2000 [13], Hickey Ju Van Emden 2001 [360]. For the latest works in interval algebra see Kreinovich 2011 [755]

3.1. Internal Representation Data and Algebra: intervals

Wang 2008 [63].

An interval can be defined: (Gioia 2006 [308] and Rokne 2001 [593]):

$$x^I = [\underline{x}_i, \overline{x}_i] = (x \in \mathbb{R} | \underline{x}_i \leq x \leq \overline{x}_i) \quad (3.3)$$

Following Revol 2009 [581] (See figure 3.1) the intervals of real numbers can be considered also as connected sets of \mathbb{R} . In this sense, these data d can be measured with an error $\pm\epsilon$. So we can have $[d - \epsilon; d + \epsilon]$

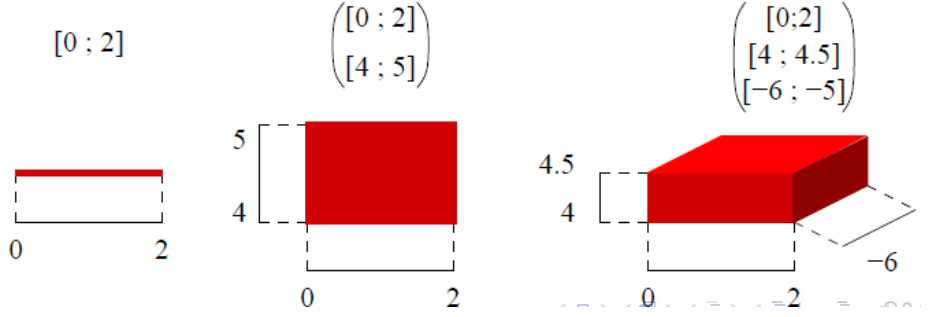


Figure 3.1: Intervals (Revol 2009 [581])

The interval valued variable can be defined in this way: we consider an interval-valued variable $[x]$ (we follow here Gioia and Lauro 2005 [310]). $X_i = [\underline{x}_i, \overline{x}_i]$ $i = 1 \dots n$. Alternatively, upper and lower bounds can be written equivalently as: $x_L \leq x_U$. A specific example could be related to the returns of n different stocks in a temporal interval t in a portfolio. So we have:

$$([\underline{x}_1, \overline{x}_1], [\underline{x}_1, \overline{x}_2], \dots, [\underline{x}_n, \overline{x}_n],) \quad (3.4)$$

Upper and the Lower bound are: Interval: $[x]$ over the base set (E, \leq) is an ordered pair, where $[x] = [x_L; x_U]$ where $x_L, x_U \in E$ are

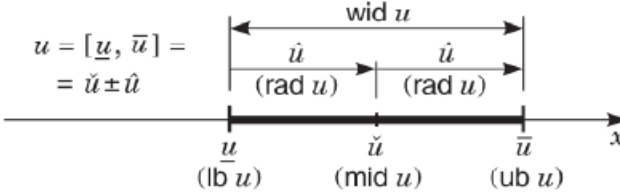
bounds such that $x_L \leq x_U$.

Another way to consider the Upper and the Lower bounds of an interval is this way:

Lower and Upper Bound: $[X]_t = [X_{t,L}, X_{t,U}]$ with $-\infty < X_{t,L} \leq X_{t,U} < \infty$

Center and Radius: $[X]_t = \langle X_{t,C}, X_{t,R} \rangle$ where $X_{t,C} = (X_{t,L} + X_{t,U})/2$ and $X_{t,R} = (X_{t,U} - X_{t,L})/2$ (See figure 3.2)

Figure 3.2: A real interval and its parameters lb (lower bound), ub (upper bound), mid (midpoint), rad (radius) and wid (width) Kulpa 2004 [435]



Considering the upper and the lower bound, an interval $[\underline{x}, \bar{x}]$, can be defined as the set of real numbers between \underline{x} and \bar{x} :

$$[\underline{x}, \bar{x}] = \{x / \underline{x} \leq x \leq x / \bar{x}\} \quad (3.5)$$

Thin intervals can be considered as: $[\underline{x}, \underline{x}] = \underline{x}$, or $[\bar{x}, \bar{x}] = \bar{x}$. Usual set theory applies: $[\underline{x}, \bar{x}] \subset [\underline{y}, \bar{y}]$. At the same time $[\underline{x}, \bar{x}] = [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} = \underline{y}; \bar{x} = \bar{y}$.

At the same time, it is possible to define the width of interval as in Rokne 2001 (see Rokne 2001 [593])

$$w(A) = \bar{x} - \underline{x} \quad (3.6)$$

at the same time we can have the absolute value of the interval:

$$|A| = \max[|\underline{x}|; |\bar{x}|] \quad (3.7)$$

a general distance between two intervals can be defined as the Hausdorff distance:

$$q(\underline{x}, \bar{x}) = \max[|\underline{x} - \underline{y}|; |\bar{x} - \bar{y}|] \quad (3.8)$$

Various contributions have been made on theoretical developments and on applications of the interval algebra¹²: Billard Diday 2000 [83].

The arithmetic is considered an extension of real arithmetic. Let \mathbf{I} be the set of closed intervals. The set of all real intervals is denoted by $I\mathfrak{R}$ and is defined as a real interval space¹³.

Operations between intervals:

Following Gioia 2006 [308] and Rokne 2001 [593], we define arithmetic operations on intervals as \bullet with the symbols $+$, $-$, Δ , $/$, \cdot .

$$[\underline{x}, \bar{x}] \bullet [\underline{y}, \bar{y}] = (x \bullet y : \underline{x} \leq \bar{x}; \underline{y} \leq \bar{y}) \quad (3.9)$$

Except the case: $[\underline{x}, \bar{x}]/[\underline{y}, \bar{y}]$ with $0 \in [\underline{y}, \bar{y}]$

The interval arithmetic can be considered as an extension of the real arithmetic. Let $[\underline{x}, \bar{x}], [\underline{y}, \bar{y}]$ in $I\mathfrak{R}$

Sum:

$$[\underline{x}, \bar{x}] + [\underline{y}, \bar{y}] = [\underline{x} + \underline{y}; \bar{x} + \bar{y}] \quad (3.10)$$

Subtraction:

$$[\underline{x}, \bar{x}] - [\underline{y}, \bar{y}] = [\underline{x} - \bar{y}; \bar{x} - \underline{y}] \quad (3.11)$$

Multiplication:

¹²At the same time, for the development of the diagrammatics of the interval algebra see the works of Kulpa in 2006 [436] and in 2001 [434]

¹³Kulpa 2001 [434]

$$[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] = [\min(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}); \max(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y})] \quad (3.12)$$

Division:

By considering $0 \notin [\underline{y}, \bar{y}]$, we have:

$$[\underline{x}, \bar{x}] / [\underline{y}, \bar{y}] = [\underline{x}, \bar{x}] \times [1/\bar{y}, 1/\underline{y}] \quad (3.13)$$

Following Revol (2009) [581] we have:

$$[\underline{x}, \bar{x}]^2 = [\min(\underline{x}^2, \bar{x}^2), \max(\underline{x}^2, \bar{x}^2)] \quad (3.14)$$

by considering $0 \notin [\underline{x}, \bar{x}]$ and $[0, \max(\underline{x}^2, \bar{x}^2)]$ in a different way.

$$1/[\underline{x}, \bar{x}] = [\min(1/\underline{x}, 1/\bar{x}), \max(1/\underline{x}, 1/\bar{x})] \quad (3.15)$$

if $0 \notin [\underline{x}, \bar{x}]$

and also:

$$\sqrt{[\underline{x}, \bar{x}]} = [\sqrt{\underline{x}}, \sqrt{\bar{x}}] \quad (3.16)$$

if $0 \leq \bar{x}$, and $[0, \sqrt{\bar{x}}]$ in a different way.

Following Revol (2009) [581], we can consider that some algebraic properties are lost. For example, the subtraction cannot be considered the inverse of the addition and at the same time division is not the inverse of the multiplication. The process of squaring is different from multiplying an interval by itself. The multiplication is sub-distributive with respect to the addition.

It is possible to extend the functions. So we have (Revol 2009 [581]) an interval extension: f of a function f satisfies $\forall x, f(x) \subset \mathbf{f}(\mathbf{x})$ and $\forall x f(x) = \mathbf{f}(\mathbf{x})$

In this case, at the same time as functions we have:

$$\exp(x) = [\exp(\underline{x}); \exp(\bar{x})] \quad (3.17)$$

$$\log(x) = [\log(\underline{x}); \log(\bar{x})] \quad (3.18)$$

if $\underline{x} \leq 0$ but $[-\inf; \log(\bar{x})]$ if $\bar{x} > 0$.

It is possible to enumerate other function examples using intervals.

By considering specifically interval data, it is interesting to note that not even the interval data comes from lower or upper bounds. In some studies there is a different choice between values in a different range (75% for example) to guarantee the absence of overgeneralization (see Chapter 1) and in that case the absence of outliers. See for example Arroyo et al. (2007) [37] in which there is specific mention.

3.1.3 Statistical methods for Interval Representations

Various methods were designed to analyze interval data, here we present a short review¹⁴. So, it is possible to define interval random variables, based on interval data (see Billard and Diday 2006 [86], Kubica Malinowski 2006 [433] and Gonz  les Rivera and Arroyo 2011 [318]). A definition of the interval random variables with an application on simulation to improve the robustness of the results is given in Betarseh and Wang 2008 [78].

There are various proposals in interval data analysis (see also Signoriello 2008 [630]).

Rodr  guez 2000 [759] and Gioia 2001 [307]) have proposed descriptive statistics for interval data, by extending the case of the single-valued or scalar data.

Gioia and Lauro 2005 [310] and Gioia 2001 [307] make a proposal of descriptive statistics based on interval data, such as mean and deviation from mean, where the author shows that the properties of the statistics considered share the same properties as the corresponding

¹⁴An updated review is present also in Diday 2008 [209]

statistics for single-valued data (Signoriello 2008) [630]).

The general approach for the advanced statistical methods is to consider statistical methods for scalar or single-valued data and extend them for the interval data¹⁵.

The extension of the Principal component analysis to interval data (from the scalar data) is developed in various scientific works: starting from radii or vertices (see Vertices Principal Components Analysis (also defined as V-PCA) in Cazes et al. 1997 [124]) and a Symbolic Objects PCA (see Lauro Palumbo 2000 [446]) to the considered mid-points (for example the Midpoint and the Radii PCA in Palumbo 2003 [543]). Irpino 2006 [393] considers an extension of the classical Principal Component Analysis to analyse time dependent interval data. Finally, Lauro Verde and Irpino in 2008 [450] review the Principal Component Analysis techniques using interval data.

Signorello 2008 [630] and Palumbo [543] report an important problem in interval algebra (related to the extension of the scalar or single-value data to interval data): "unfortunately, the interval algebra was born in the field of error-theory where intervals are very small, but this is no longer true for Statistical Interval. First of all the so-called wrapping effect leads to wider intervals than they actually should be. This effect induces a distinction between "interval of solutions" and the "interval solutions" .

It is not possible here to present all the works that extend the scalar or single-valued data statistical techniques, for an updated review see Diday, Noirhomme (2008) [218].

¹⁵Signoriello 2008 [630]

3.1.4 Stochastic Processes and Time Series of Interval-Valued Representations

There is a growing literature in Interval Data and Interval Time Series (ITS). In fact, in recent years there has been a literature that considers more in depth the methods of forecasting using these types of representations. A first contribution is in Teles and Brito (2005) [654], a useful paper is that of Cheung who introduces the forecasting of daily highs and lows [133]¹⁶, at the same time Maia, De Carvalho and Ludermir in 2006 and in 2008 [477] [476] introduce various methods in forecasting intervals, in particular hybrid methods. Different forecasting methods using interval data are considered in Arroyo, Muñoz San Roque, Maté, and Sarabía (2007) [37], in particular exponential smoothing, Arroyo Gonzáles Rivera and Maté 2010 [41] on VAR and KNN methods applied in interval forecasting¹⁷. Maté and García Ascanio (2010) [494] compare different forecasting methods (VAR and Neural Networks) using energy data whereas Arroyo Espínola and Maté (2010) [36] consider financial data in the method comparison.

Another very important paper that develops methods for interval time series (ITS) is: Han, Hong and Wang (2009) [336]. An application on exchange rates is in Han Hong Lai Wang 2008 [337] and He Hong an Wang 2011 [351] with an application on crude oil prices. Forecasting combinations with interval data is the topic of the work of Salish and Rodrigues 2010 [603]¹⁸.

¹⁶This paper is particularly important because it introduces some useful methods used in forecasting intervals. See also Chou who introduces CARR models for volatility modelling (Chou 2005 [135]), Rogers and Satchell 1991 [592] Spurgin and Schneeweis 1999 [638] and Parkinson 1980 [548]

¹⁷The distance for interval data are presented in Arroyo Maté 2006 [43]

¹⁸The author in another work in 2010 and 2011 found evidence of nonlinearity in financial interval time series (ITS) [602] and [591]

Summary Results: Intervals
The simplest Internal Representation.
Assume data to be a Mixture of Uniforms.
Interval Algebra as extension of the rules of Arithmetic Algebra.
Interval Algebra can be considered to be the foundation of the other Representations.
Upper, Lower Bound, Centres and Radii can be descriptors, and can be considered over time with Attribute Time Series.

Chapter 4

Foundations of Boxplots and Histograms Data Representations

In this chapter we deal with density valued data methods and their algebra. We have seen in the Third Chapter representation characterized only by two relevant measures (the upper and the lower bound, or the centre and the radius). We have seen that this representation could be very important if we are interested in extreme values of an aggregate data in preserving some of the original data structure or the initial distribution (in practice, we are assuming that data follows a uniform distribution). In any event, we shall see in this chapter and in Chapter 5 that this representation could be improved by considering other types of representations like the boxplots or the histograms, and finally in Chapter 5 the densities.

At the same time in this chapter we will make for the first time the distinction between original data and model data to take into account the measurement model that could be determined by various factors. Finally in Chapter 7 we will show how to estimate the coefficients of

these data.

4.1 Internal Representation Data and Algebra: Boxplots, Histograms and Models

4.1.1 Quantile Data and Algebra

In particular there was some debate on the interval arithmetics because there were some relevant cases in which interval is not so useful in representing some variables. Williamson 1989 [701] speaks of a "number of techniques which determine limited information about the distribution of functions of random variables".

At the same time, following Williamson 1989 [701] the general aim is to provide information on the distribution of the variable for the internal representations, and in this respect it is necessary to consider some types of generalizations of the interval algebra we have already looked at. Various attempts have been made in this sense, Williamson 1989 [701], in practice cites two approaches: Triplex Arithmetic and the Quantile Arithmetic.

The problem related to Ecker and Ratschek (1972) [245] as explained in Williamson (1989) [701] is that they "have considered intervals probabilistically in an attempt to understand the phenomena of subdistributivity and inclusion monotonicity. They also suggested a joint representation of distributions and intervals and studied some properties of Dempster quantile arithmetic which we examine below". At the same time Williamson (1989) [701] says, about the approach of Ahmad 1975 [9]: "Ahmad is supposed by Moore 1979 [514] to have looked at the arithmetic of probability distributions from the point of view of interval arithmetic. However Ahmad's paper is solely con-

4.1. Internal Representation Data and Algebra: Boxplots, Histograms and Models

cerned with nonparametric estimators of probability densities and he has nothing to say about probabilistic arithmetic”.

So the two different frameworks considered and presented sequentially (See figure 4.1 and figure 4.2), here are the Triplex Arithmetic and the Quantile Arithmetic.

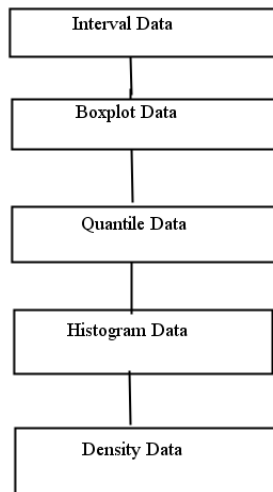


Figure 4.1: Comparing Internal Representations

Triplex Arithmetic and Quantile Arithmetic

We start from the problem we have considered in Chapter 3:

$$I_{\mathfrak{R}} = ([\underline{x}, \bar{x}] \mid \underline{x} \leq \bar{x}, \underline{x}, \bar{x} \in \mathfrak{R}) \quad (4.1)$$

Two intervals could be considered binary operations, as c for example, $Z^I = X^I Y^I$ that could be given by:

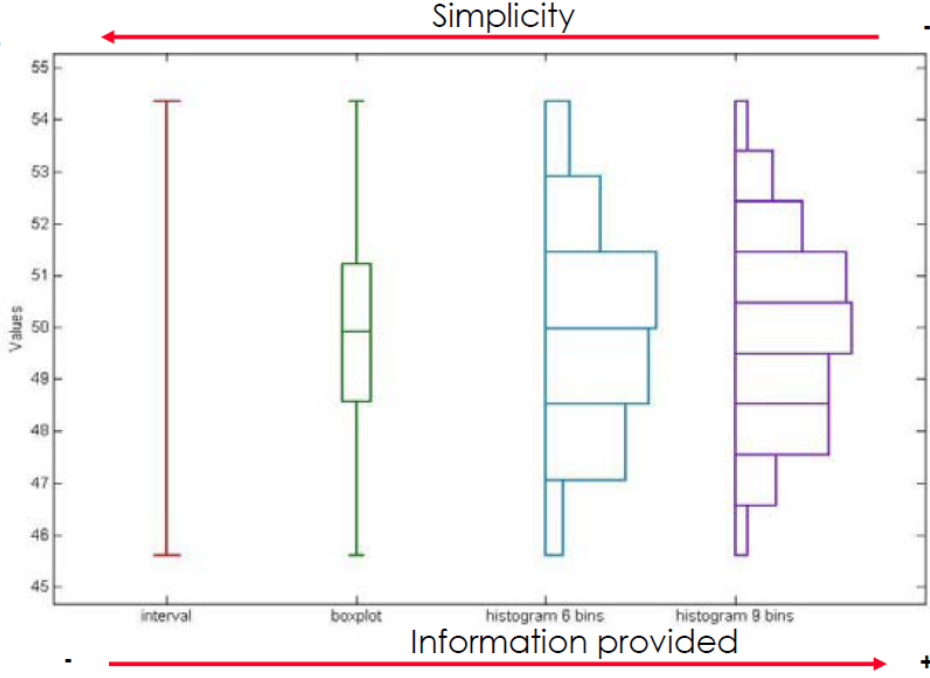


Figure 4.2: Comparing Complex Internal Representations: Sampled data from a $N(50,2)$ summarized by symbolic variables — González-Rivera G. Carlos Maté (2007) [315]

$$Z^I = [\underline{z}, \bar{z}] = (xy \mid x \in X, y \in Y) \quad (4.2)$$

Sometimes it is useful to consider some generalizations regarding this algebra. Following Williamson 1989 [701] in particular there were two different developments: the Triplex Arithmetics and Quantile Arithmetics. The Triplex Arithmetics (see Cole and Morrison 1982 [149] Apostolatos et al. 1968 [24] and Nickel 1969 [538]) consider, in

addition to the extreme values of interval, a main value or the most probable one. In particular: a triplex number X^t can be defined as an ordered triple $[a, c, b]$ in which a could be considered the lower bound, the c is considered the most probable value (or the main value) and the b is the upper bound.

$$[\underline{x}, \tilde{x}, \bar{x}] \quad (4.3)$$

It is interesting to note that if we only consider \underline{x} and \bar{x} we can use mainly the rules of interval arithmetics (see Williamson 1989 [701]). The main objective of the Triplex Arithmetic is to give an answer to the limits of the interval arithmetic by providing some information on the underlying distribution.

Following Williamson 1989 [701] the Quantile Arithmetics results are related to the work of Dempster (see Dempster 1974 [191] and 1980 [192] and Dempster and Papagakipapoulis 1980 [194]).

It is necessary to consider a random variable X with a density f_X and a distribution function F_X . It is possible to define the quantile number X^Q , which represents X by the approximation F_{X^Q} to f_X

$$f_{X^Q}(x) = \begin{cases} \alpha & \text{if } x = F_x^{-1}(\alpha) \\ 1 - 2\alpha, & \text{if } x = F_x^{-1}(\frac{1}{2}) \\ \alpha & \text{if } x = F_x^{-1}(1 - \alpha) \\ 0 & \text{otherwise} \end{cases} \quad (4.4)$$

It is important to note that f_X is a density where f_{X^Q} is a discrete frequency function¹. It is possible that two quantile numbers X^Q and Y^Q are combined in the way that $Z^Q = X^Q \square Y^Q$ by the rule:

¹Williamson 1989 [701]

$$f_{Z^q}(z) = \begin{cases} f_{X^q}(x)f_{Y^q}(y) & \text{for } z = x \square y \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

However Williamson 1989 [701] concludes that quantile arithmetics comes to an underevaluation of the intervals, whereas the interval arithmetics tends to overestimate the spread.

Boxplot Data

A special case related to the boxplots is proposed by Arroyo and Maté in 2006 [44] and it is related to Boxplot Data Analysis. In particular in this case there is the use of interval algebra in boxplot data analysis of the internal representations.

This method comes to a different data representation that could be more useful in understanding the data structures. In particular:

$$Z(u) = m_u, q_u, Me_u, Q_u, M_u \quad (4.6)$$

With:

$$-inf \leq m_u \leq q_u \leq Me_u \leq Q_u \leq M_u \leq inf \quad (4.7)$$

In a first paper, Arroyo and Maté 2006 [44] describe the statistical methods that could be applied to the boxplot data. In a second paper, Arroyo Maté and Munoz 2006 [42] develop methods for boxplot variables, as for example hierarchical clustering.

Boxplot Time Series (BoTS)

Arroyo Maté and Muñoz A. (2006) [42] use Boxplot Time Series (BoTS figure figure 4.3), while the same time series are discussed in Maté, Arroyo (2006). [493]. In a recent paper Arroyo (2010) [33] introduces a new tool similar to the boxplot: the Candlestick time series CTS (in

4.1. Internal Representation Data and Algebra: Boxplots, Histograms and Models

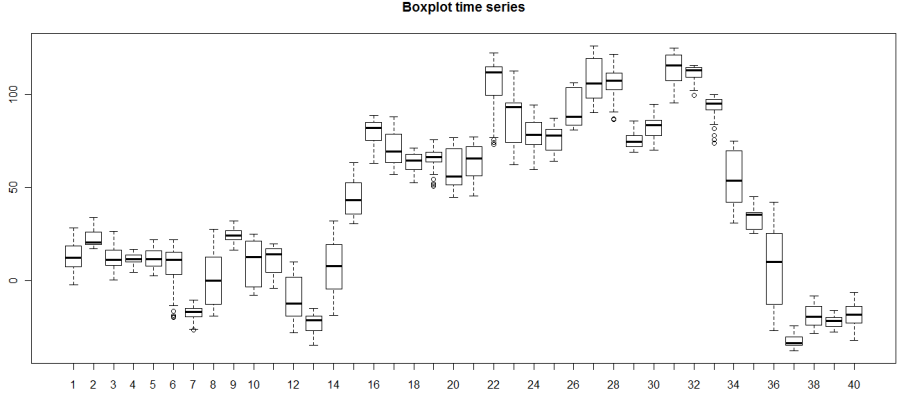


Figure 4.3: Boxplot time series (BoTS)

finance, they are the Japanese Candlesticks figure 4.4) predicted with locally weighted learning methods (see also Arroyo Bomze 2010 [35])

4.1.2 Histogram Data and Algebra

Histograms are an alternative type of data, that offer an answer to the same problem of the boxplot data (figure 4.5). The difference is straightforward, by defining the optimal number of the bins the histograms represents the bumps of the data, whereas the boxplots typically do not.

For the Histogram algebra, see the work of Williamson 1989 [701] and also Billard and Diday 2010 [88], about the histogram algebra see Colombo and Jaarsma 1980 [150]. An explanation of the histogram algebra is given in Gonzáles-Rívera and Maté 2007 [315] An interesting application of the theoretical methods is in Arroyo et al. (2011) [38]. The authors proposed also an algorithm for the conversion between

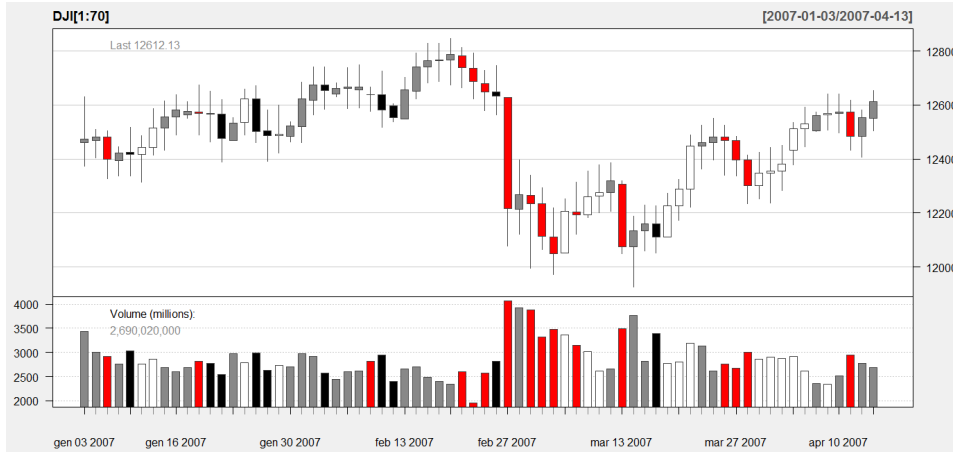


Figure 4.4: Candlestick time series (CTS)

histograms and quantiles.

An alternative approach to the histogram algebra is given in Carreras and Hermenegildo 2000 [121], where it is possible also to consider Gupta and Santini 2000 [330], and Luo Kao Pang 2003 [473] on "visualizing spatial distribution data sets".

Another different approach is given in the project AIDA (A.A.V.V. [727]) in which the analysis and the operations are performed bin by bin. Following Signoriello (2008) [630]: there are relevant cases in which data are collected and can be faithfully represented by using some frequency distributions.

Assuming X as a variable numerical and continuous, we can observe various different values x_i . We are specifically interested in its variation. The values can be regrouped in a smaller number H of consecutive and disjoint bins I_h . These values give the internal variation of the representation requested.

By considering the number of data n_h belonging to each I_h we ob-

4.1. Internal Representation Data and Algebra: Boxplots, Histograms and Models

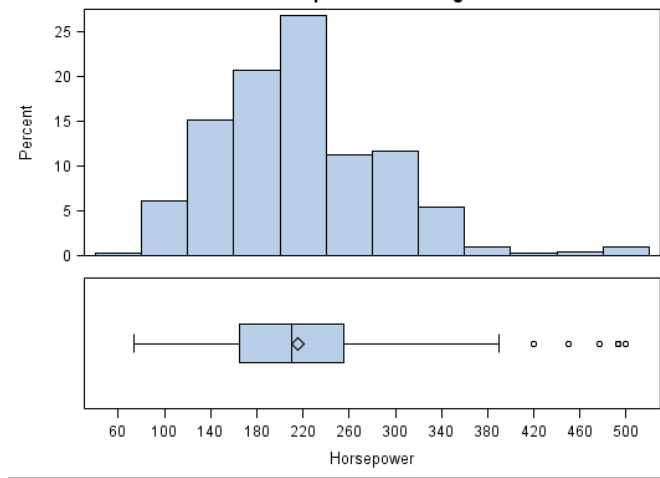


Figure 4.5: Differences between Histogram and Boxplot Data. Source SAS 9.2 Support Documentation

tain the frequency distribution of the variable X . So in this case, we can consider the histogram as a specific internal representation of a variable X

Histogram data offers much literature on symbolic data analysis methods. Signoriello in 2008 [630] presents a definition of the histogram data as follows:

We assume X being a continuous variable defined on the finite support $S = [\underline{x}; \bar{x}]$, in which \underline{x} and \bar{x} are the lower and upper bounds of the domain of X .

In this case, the variable X can be considered as divided and represented on a set of adjacent intervals, that are defined as the histogram bins $I_1, \dots, I_h, \dots, I_H$, where $I_h = (\underline{x}_h; \bar{x}_h)$.

Given a number n of observations on the variable X , each semi-open interval, I_h , could be associated with a random variable equal to

$\phi(I_h) = \sum_{u=1}^N \phi_{x_u}(I_h)$ where $\phi_{x_u}(I_h) = 1$ if $x_u \in I_h$ and 0 otherwise.

At the same time, it is possible to associate to I_h an empirical distribution $\pi_h = \phi(I_h)/N$.

In that sense a histogram X (figure 4.6) can be defined as an internal representation in which each pair $(I_h; \pi_h)$ (for $h = 1, \dots, H$) is defined both by a vertical bar, with base interval I_h along the horizontal axis and the area proportional to π_h .²

Consider E as a set of n empirical distributions $X(i)$ ($i = 1, \dots, n$).

It is important to note that, compared with the interval data, which is usually representing a uniform distribution, histogram data, given the X variable, the i -th can represent an empirical distribution defined as a set of H ordered pairs $X(i) = (I_h, \pi_h)$ as:

$$I_{hi} \equiv [\underline{x}_{hi}, \bar{x}_{hi}] \quad \underline{x}_{hi} \leq \bar{x}_{hi} \in \mathfrak{R},$$

$$\bigcup_{h=1, \dots, H} I_{hi} = [\min_{h=1, \dots, H} \{\underline{x}_{hi}\}, \max_{h=1, \dots, H} \{\bar{x}_{hi}\}],$$

$$\pi_h \geq 0,$$

$$\sum_{h=1, \dots, H} \pi_h = 1. \tag{4.8}$$

The Histogram data shows (Gonzales Rivera and Maté 2007 [315]):

$$c(h) = \sum_{i=1}^p \frac{\underline{I}_i + \bar{I}_i}{2} \pi_i \tag{4.9}$$

The problem with the histogram is the choice of the number of bins. A possible solution could be the answer in Sturges 1926 [645]:

²Signoriello 2008 [630]

4.1. Internal Representation Data and Algebra: Boxplots, Histograms and Models

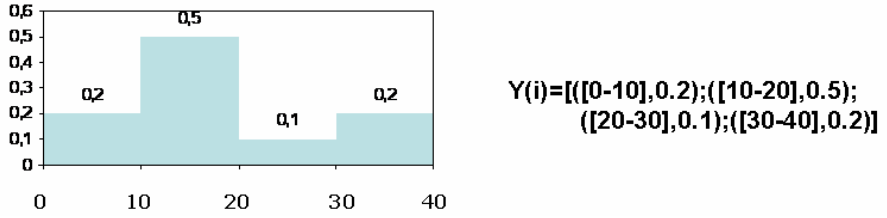


Figure 4.6: Histogram Data

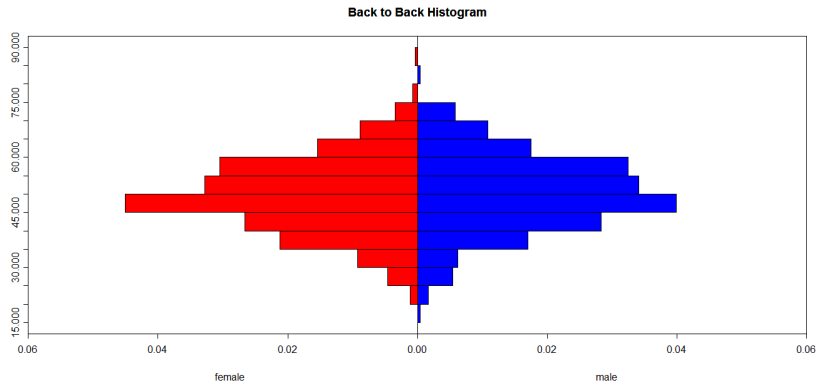


Figure 4.7: Back to Back Histograms

$$\hat{p} = 1 + \log_2 n \quad (4.10)$$

and also

$$\hat{h} = \frac{\max(x) - \min(x)}{1 + \log_2 n} \quad (4.11)$$

Following Signoriello 2008 [630] these data types show a two-dimensional

representation, with an interval division in intervals (between the breaks) and the densities vertically.

Histogram data can be considered a special case of the internal representations that could be divided into intervals where each provide information on the relative frequency.

It is possible to work with intervals of different weights given by the respective frequency. In particular Colombo and Jaarsma (1980) [150] proposed a histogram arithmetic as follows:

Given two histograms figure 4.7, $Y_A = (I_{Ah}, \pi_{Ah})$ with $h = 1, \dots, n$ and $Y_B = (I_{Bh'}, \pi_{Bh'})$ with $h' = 1, \dots, m$ both representing a pair of independent random variables A and B , and \square being some arithmetic operator in $\{+, -, \times, \div\}$, $C = A \square B$ can be approximated by the unsorted histogram $Y_C = (I_{Ck}, \pi_{Ck})$ with $k = 1, \dots, n \cdot m$, where

$$\underline{x}_{C(h-1)m+h'} = \min \{ \underline{x}_{Ah} \square \underline{x}_{Bh'}, \bar{x}_{Ah} \square \underline{x}_{Bh'}, \underline{x}_{Ah} \square \bar{x}_{Bh'}, \bar{x}_{Ah} \square \bar{x}_{Bh'} \}, \quad (4.12)$$

$$\bar{x}_{C(h-1)m+h'} = \max \{ \underline{x}_{Ah} \square \underline{x}_{Bh'}, \bar{x}_{Ah} \square \underline{x}_{Bh'}, \underline{x}_{Ah} \square \bar{x}_{Bh'}, \bar{x}_{Ah} \square \bar{x}_{Bh'} \} \quad (4.13)$$

$$\pi_{C(h-1)m+h'} = \pi_{Ah} \square \pi_{Bh'} \quad (4.14)$$

There are some disadvantages in using the histogram algebra (Signoriello 2008 [630] and Colombo and Jaarsma [150]), in particular:

1. Some intervals, as a result of the computations, may overlap³
2. It is important to note that by performing a series of arithmetic operations on a number of histograms the resulting histogram is

³Signoriello 2008 [630]

4.1. Internal Representation Data and Algebra: Boxplots, Histograms and Models

expected to have a higher number of intervals than the starting histograms.

3. It is possible to compact the unsorted histograms obtained after each operation in order to avoid an enormous final number of intervals.
4. The interval arithmetics is the foundation of the rules of histogram arithmetics. In the same way, interval arithmetics is based on classical arithmetics⁴.

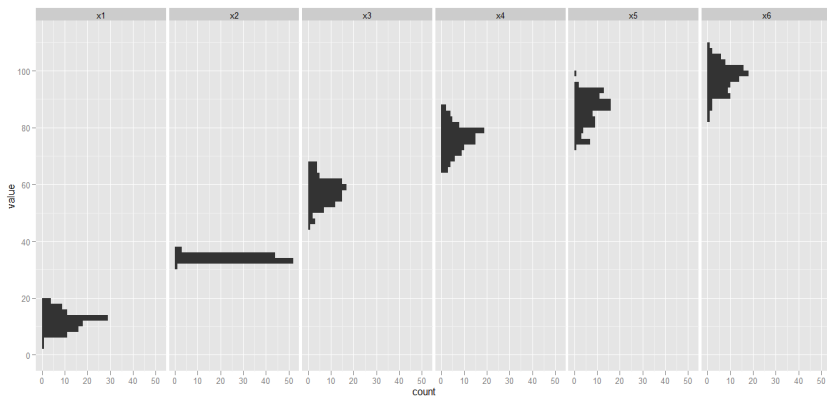


Figure 4.8: Histogram Time Series (HTS)

⁴Signoriello 2008 [630]

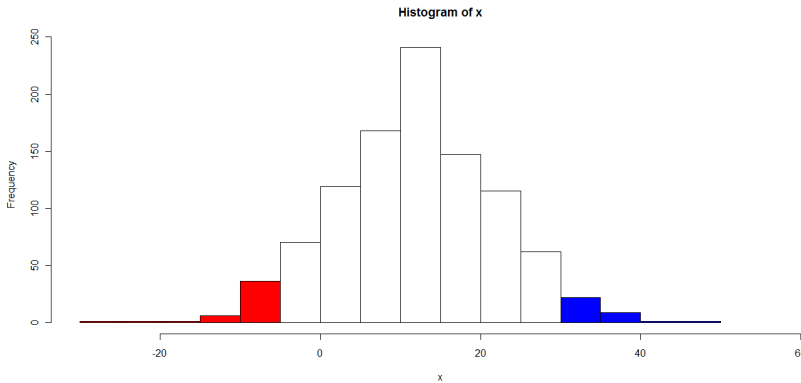


Figure 4.9: Clipping Histograms (Risk Visualization)

4.2 Statistical Methods Involving Boxplots and Histograms valued data

For every type of internal representation of the data, intervals, histograms etc., there are various proposals about statistical methods. Various reviews can be considered, for example Diday and Noirhomme 2008 [218] and Signoriello 2008 [630].

In this paragraph we will review the statistical techniques using Internal Representations, in particular Histogram Data.

Colombo and Jaarsma 1980 [150] describe the histogram rules that could be considered to define statistical methods for histogram data.

Works on Principal Component Analysis of Histogram Data are from: Rodríguez Diday and Winsberg 2000 [587] Nagabhushan and Pradeep Kumar 2007 [527] and Diday 2011 [213].

Approaches to Histogram data regression are in Verde and Irpino 2011 [677] Dias and Brito 2011 [197] and Wang Guan Wu 2011

4.2.1 Histogram Stochastic Processes and Histogram Time Series (HTS)

Very recent and relevant is the growth in literature of the symbolic forecasting methods using histogram time series (HTS figure 4.8). Arroyo and Maté 2009 analyse in depth interval data forecasting methods using KNN K-Nearest Neighbour methods [44]. González Rivera and Arroyo 2011 [319] and [318] define the concept of the histogram random variable and stochastic process. The stochastic process can be defined as: A histogram-valued stochastic process is a collection of histogram random variables that are indexed by time. A histogram time series (HTS) is the realization of a histogram valued stochastic process. A histogram valued stochastic process can be defined as weak stationary if every interval are weakly stationary processes. An important definition is the barycentric histogram that minimizes the distances with other histograms and is obtained by optimization. The authors derive the empirical autocorrelation with respect to the barycenter (See Gonzales Rivera and Arroyo 2011 [319] and [318]). An important way to analyse histogram time series (HTS) is to consider quantile intervals in the histograms (figure 4.9), in particular Gonzales Rivera and Arroyo 2011 [318] show the usefulness of this type of analysis.

4.3 Internal Representations Models

Until now, we have assumed that initial data are not affected by the existing error. There are important cases in which this assumption is untenable. In these cases, it is necessary to consider the existence of the noise in the data, and "model" the data accordingly. According to

Signoriello (2008) [630], and also Drago, Lauro and Scepi 2009 [233], sometimes it can be very useful to model the shape of the initial data (say, a histogram or later the density data) to extract the relevant information from our data. In this case, we are taking into account the errors present in data in a wide sense (for example missing value, or other phenomena)⁵. We define data model as specific data, in which we have eliminated the measurement error.

It is important to note real models do not exist but any model can be an approximation of the reality. A model is good if it is useful as approximation. In this sense, only models that could be validated by data can be useful.

Tukey in a work of 1977 [670] and Caussinus in 1986 [123] allow in what sense it is possible to interpret the statistical methods (both confirmatory and exploratory) by approximating initial data as:

$$Data = Structural \ Part + Noise \quad (4.15)$$

Where D are the data, S is the structural part and N is the noise. This idea is relevant for example in the time series analysis, in which we want to extract the signal from the noise.

The idea proposed by Signoriello 2008 [630] and Drago, Lauro and Scepi 2009 [233] is that of specifically transforming the histogram data by using an approximation function to control the error (or the noise E) deriving from empirical data figure 4.10.

In this case, if we assume the structural part as our model we can reformulate the empirical data as:

$$Data = Model + Error \quad (4.16)$$

Here M is the model used and E the error. In this case the data are specifically obtained by the approximated function in the modelling

⁵Signoriello 2008 [630]

process and we obtain the model coefficients and an index of goodness of fit.

The general problem is to find some functions (the models) that optimally approximate the initial data (histogram, for example) as polynomial models, splines, or B-splines. Each different model can be characterised by a goodness of fit (an evaluation index of the quality of the representation of the initial data). Later we will express these data as a finite mixture model in which we obtain the same result to extract the relevant information from the original data.

So for each i -observation we obtain a specific n -function. It is possible to represent the coefficients as in the figure. It is important to note that the original histograms or the original data are substituted by their coefficients.

4.4 The Data Choice

Simple examples of conversion between different data types are in Arroyo et al. 2011 [38] in which the authors propose an algorithm to transform initial histogram data into quantiles.

4.4.1 The Optimal Data Choice

The choice of the best interval representation data is attributed to the analyst and to the specific analysis (so they can be considered a choice to perform a priori). Where there is not a specific a priori preferred data type there are two fundamental factors that need to be considered:

1. A genuine data type (in the sense there is no possibility of choice because the data needs to be considered).

FOUNDATIONS OF BOXPLOTS AND HISTOGRAMS DATA REPRESENTATIONS



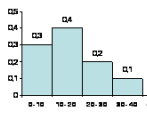
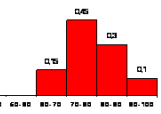
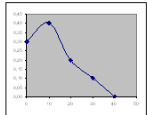
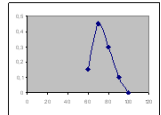
		Variable Y_1	Variable Y_2
Interval data			
Histogram data			
"Model data"			

Figure 4.10: Different types of Symbolic Data (Signoriello 2008 [630])

2. The distance from a uniform distribution for the original data. The more similar the original data is to a uniform the better the approximation to an interval.
3. The usefulness of prediction of some specific values (for example in Environmetrics, typically the maxima or the minima for a specific temporal interval).

Simulations tell us that the loss of information is related to the distance from the uniform, and from the number of the primary individuals in the aggregation interval.

4.4.2 Conversions between Data

It is possible to convert the data, when the internal representation become from a summary of huge sets of data, see Diday Esposito 2003 [216]. In fact, from the database queries we can extract different categories of descriptive variables. At the same time, a specific representation comes from the use of a multidimensional specific technique that perform a partitioning in the original massive data set. In all these cases, it is possible to change the internal representation of the data; where the data is specifically native this practice is straightforward, see for example Diday and Noirhomme 2008 [218]. There is a unique problem to consider in the choice, that one choice or another one could determine a specific loss of the variability of the data [212]. The problem is clear when it is necessary to transform the original internal representation into classical data here there is a loss of information (due to the fact that the classical data does not consider variation).

So when the data shares a strong internal variability and heterogeneity, they need to be represented as internal representations over time or over space by taking into account this variability.

Summary Results: Histograms
An Internal Representation preserving complex patterns of intra-data variation.
Histogram Algebra can be built on Interval Algebra.
It is possible to estimate the relevant coefficients to obtain the Data Models (Histograms can be parameterized to obtain Data Models).

Chapter 5

Foundations of Density Valued Data: Representations

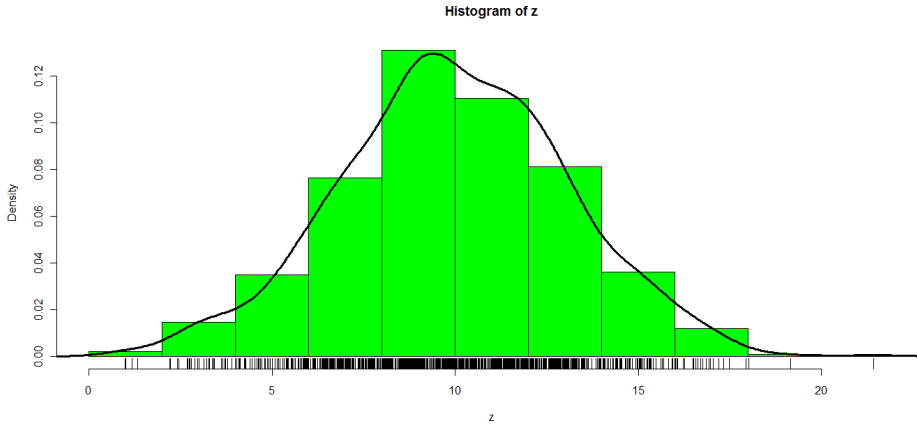
In general, it is possible to use other statistical methods to represent some data. In this way, the histogram or the boxplot (both seen in Chapter 4) are a type of the representation that could be possible to be considered. In literature it is sometimes suggested to use kernel density estimators instead of histograms for the smoothness and the bin placement (see a discussion in Di Nardo Tobias 2001 [222] but also in Gelman 2009 [293] for a discussion of the topic), in fact both are nonparametric methods which do not impose any type of specific parametric form. In the histogram case¹, in practice, bins can determine the shape of the density and the discontinuities and sometimes it is useful to explore also this aspect by removing the discontinuity². So,

¹There are at the same time proposals for different histogram types, see for example: [195]

²The histograms by definition are not smooth. The histogram shape depends on the bin width and the end point of the same bins. In this way in constructing

a more reasonable choice could be directed in to exploring underlying data structure and avoid the bin choice. Also the choice to use models (Chapter 7) and a coefficient estimation, using for example some mixtures, could be addressed to eliminate the error and in defining the underlying data structure.

Figure 5.1: Kernel density estimation, histogram and rugplot on simulated data



5.1 Kernel Density Estimators

The simplest non parametric density estimator is the histogram: in fact we can approximate the density by the fraction of points that fall in the bin. So:

histograms we have to consider the width of the bins and the end points of the bins

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n 1_{x_i} \quad (5.1)$$

Where x_i is in the same bin as x (Racine 2008 [569]) and $1(A)$ is an indicator function with value 1 if A is true, zero otherwise.

In this sense the histogram construction³, can start from an origin x_0 and a bin width h . The bins can be considered as $[x_0 + mh, x_0 + (m+1)h]$ both for positive and negative integers⁴.

It is possible to represent a histogram by smoothing their bins (figure 5.1 and figure 5.3 and obtaining a single continued function $f(x)$. So the kernel density estimator is⁵:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (5.2)$$

where $K(z)$ is a Kernel, a symmetric weight function and h is a smoothing parameter defined as a bandwidth⁶. The bandwidth choice in this case is very important⁷. In fact from the h depends the level of smoothness of the density. For the bandwidth selection problem see Turlach 1993 [669], Chiu 1991 [139], Hart Vieu 1990 [342]. In general a presentation and a review of a nonparametric approach is to be found in Li and Racine 2005 [465].

K can be a gaussian function with mean zero and variance 1. The Kernel is a non-negative and real-valued function $K(z)$ satisfying:

³In particular (Milani 2008 [504]) for a review of the graphical exploratory techniques

⁴Li Racine 2007 [465] and also Racine 2008 [569]

⁵Racine 2008 [569] and Pagan Ullah (1999) [542]

⁶Racine 2008 [569] states that this estimator could be defined as the Rosenblatt–Parzen estimator: see Rosenblatt (1956) [595] and Parzen (1962) [549]. Where the x_i are time series or dependent data: see Hansen 2009 [338] Wand and Jones 2005 [687] and Harvey and Oryshchenko 2010 [345]

⁷Katkovnik Shimulevich 2000 [419] Silverman 1978 [633] and Raudys 1991 [578]

$$\int K(z)dz = 1, \int zK(z)dz = 0, \int z^2K(z) = k_2 < \infty \quad (5.3)$$

with the lower and upper limits of integration being $-\infty$ and $+\infty$. $K(z)$ can be different Kernel functions figure 5.2: uniform, triangle, epanechnikov, quartic (biweight), tricube (triweight), gaussian and cosine. A popular kernel choice is the gaussian one⁸:

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (5.4)$$

Another frequent choice is the Epanechnikov Kernel, optimal in a minimum variance sense (see Epanechnikov 1969 [256]).

$$K(z) = \frac{3}{4}(1 - z^2) \mathbf{1}_{\{|z| \leq 1\}} \quad (5.5)$$

Another important choice could be the kernel triangular:

$$K(z) = (1 - |z|) \mathbf{1}_{\{|z| \leq 1\}} \quad (5.6)$$

In any case, for the kernel triangular and others the loss of efficiency is not relevant (see Wand Jones 1995 [687]).

5.2 Properties of the Kernel Density Estimators

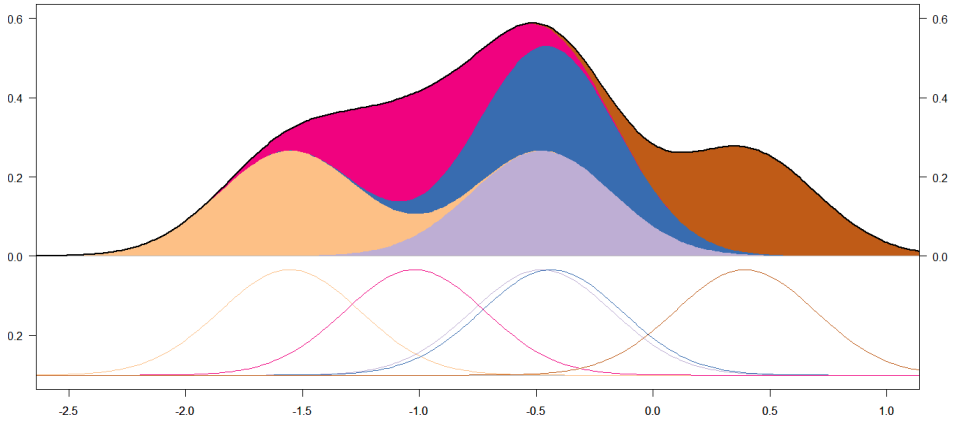
Hansen (2009) [338]⁹ also shows that if $K(z)$ is non negative it is possible to show that $\hat{f}(x) \geq 0$. At the same time it is possible to

⁸See Katkovnik Shmulevich 2000 [419]

⁹See also Baldini Figini Giudici 2006 [55]

5.2. Properties of the Kernel Density Estimators

Figure 5.2: Kernel density estimation: illustration of the kernels (Francois 2011 [280])



compute the numerical moments of the density $\hat{f}(x)$ so the mean of the estimated density will be the sample mean of X , the set of data:

$$\int_{-\infty}^{\infty} x \hat{f}(x) dx = \frac{1}{n} \sum_{i=1}^n X \quad (5.7)$$

The second moment will be:

$$\int_{-\infty}^{\infty} x^2 \hat{f}(x) dx = \frac{1}{n} \sum_{i=1}^n X^2 + h^2 k_2(k) \quad (5.8)$$

So the variance for the density \hat{f} will be:

$$\int_{-\infty}^{\infty} x^2 \hat{f}(x) dx - \left(\int_{-\infty}^{\infty} x \hat{f}(x) dx \right)^2 = \hat{\sigma}^2 + h^2 k_2(k) \quad (5.9)$$

Where in this sense the $\hat{\sigma}^2$ is the sample variance. At the same time the density estimation increases the sample variance by the $h^2 k_2(k)$

(see Hansen 2009 [338]).

At that point it is useful to consider the pointwise mean square error (MSE) criterion (following Racine 2008 [569]) that is used for analysing the properties of many kernel methods.

So it is necessary, as well, to derive the bias and the variance for $\hat{f}(x)$ to have an expression for the MSE. We can obtain the approximate bias for $\hat{f}(x)$:

$$\text{bias } \hat{f}(x) \approx \frac{h^2}{2} f''(x) k_2 \quad (5.10)$$

And at the same time the approximate variance will be:

$$\text{var } \hat{f}(x) = \frac{f(x)}{nh} \int K^2(z) dz \quad (5.11)$$

Pagan and Ullah (1999) [542] and Li and Racine (2007) [465] show the detailed derivation of the results.

It is interesting to note that bandwidth h determines the bias and the variance, with h decreasing the bias falls and the variance is higher¹⁰.

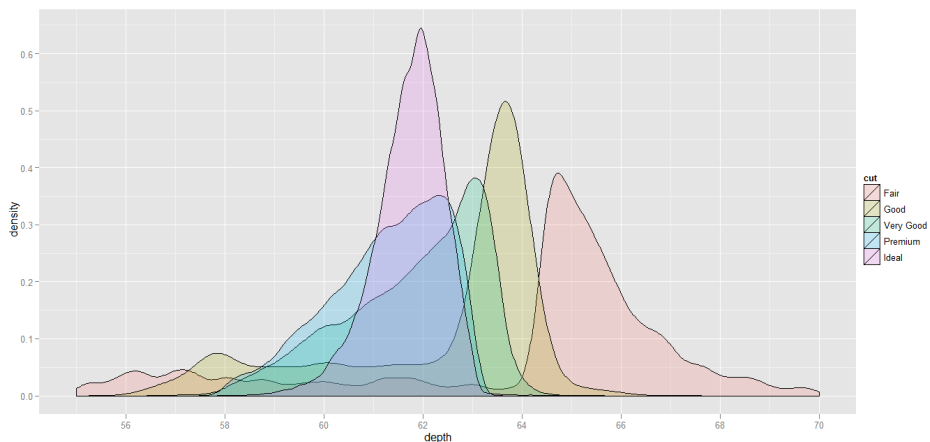
Integrated Mean Square Error (IMSE) aggregates the MSE over the entire density and could be considered a global error measure.

5.3 The Bandwidth choice

A more relevant choice is on the bandwidth h . It is important to note that the parameter h is controlling the smooth of the function: a higher smooth means a higher smooth level whereas a lower smooth means the contrary. Various methods have been proposed in literature in this sense. Racine 2008 [569] states that there are four categories in bandwidth selection:

¹⁰Racine 2008 [569]

Figure 5.3: Overlapped Kernel density estimations [793]



1. rule-of-thumb reference
2. plug-in methods
3. cross-validation methods
4. bootstrap methods

In that case, by following Racine 2008 [569] we can explore the different groups of methods. In the first case a standard family of distributions to obtain a value for the unknown constant $\int f''(z)^2 dz$ is hypothesized. So the Gaussian Kernel is used and the result will be: $1.06\sigma n^{1/5}$ as a rule-of-thumb. The sample standard deviation $\hat{\sigma}$ is used (Racine 2008 [569]).

Another known method family is the Sheather Jones method (1991), as defined in the Plug-in methods class (see Sheather Jones 1991 [624]).

The method of the Least Squares Cross-Validation is based on the idea that the bandwidth selected needs to minimize the IMSE of the

estimate.

In the methods of Likelihood Cross-Validation it is necessary to choose h to maximize the log likelihood.

Faraway and Jhun (1990) [262] have proposed a different approach based on bootstrap, in which the selection of the h is in estimating the IMSE and minimizing them over all bandwidths.

In addition, we consider data-driven methods, that are not a guarantee of good results everytime¹¹.

In the data visualization part (Chapter 6) we will use the Sheather-Jones criteria that defines the optimal h in a data-driven choice (see Kampstra 2008 [416]).

5.4 Density Algebra using Functional Data Analysis

By using Functional Data Analysis it is possible to transform the density into a functional data. The density could be considered a function and so in that sense specifically used for some operations. So it is possible to consider the beanplots and transform the densities into functional data¹²

An important result in this topic is provided by Zhang 2007 [717] Zhang and Muller 2010 [720] and by Kneip Utikal 2001 [426] and Jones 1992 [412]. Ramsay and Silverman [574] perform a principal components analysis of the log densities.

¹¹Racine 2008 for an explanation. [569]

¹²See for example Delaigle and Hall [188] who consider the approach of densities as functional data. Another approach in this sense is provided by Delicado (2010) [189] in the approach of dimensionality reduction of densities as functional data.

5.5 Density Algebra using Histogram Algebra

Density Algebra is a way to obtain for example the density mean between a group of densities by using the histogram algebra. In this sense we transform the original density data into Histograms, with regards to the method for the construction of the optimal histogram see: Scott 1979 [617], Sheather Jones (1991) [624] and Wand (1995) [686]. The number of bins needs to be chosen for the histograms to use in the analysis. So if the histograms have the same number of bins they can be computed, if not it is necessary to compute an average. After the operation of transformations we can consider the related algebraic operations between histogram data (see Colombo and Jaarsma 1980 [150]). An explanation in terms of histogram data is given by Gonz  les Rivera and Mat   2007 [315]

5.6 Density Trace and Data Heterogeneity

Sometimes a relevant assumption on data is that they represent some patterns of heterogeneity. Mixtures are useful to understand interesting patterns in data that could be detected and exploited over time t . In particular (see Ingrassia Greselin Morlini 2008 [392]) we assume a set of data Ξ that could be constituted by g different subgroups at a specific time t , so we have $\Xi = \Xi_1 \cup \dots \cup \Xi_g$ every time t . The elements for each t are mixed proportionally $\alpha_1, \dots, \alpha_g$. They can be, for example, prices of houses in a set of data Ξ in different zones. In that case there is a homogeneous quantity X intra groups and a heterogeneous one between. The random variable X can have a probability distribution different for each group, and we assume that the distribution of prob-

Data: A set of kernel density estimations

Result: The sum in a form of histogram

begin

for $i \in I$ **do**

 | Transform the density in histograms i

 | Define the optimal number of bins k

end

 Is it possible to compute the objects?

if *the objects cannot be computed* **then**

 | change the number of bins structure considering an
 | average of the bin

end

 Compute the sum of the histograms

end

Algorithm 1: Sum of Kernel Densities by Histograms

5.7. *Conversions between Density Data and other types of data*

ability intra-group are among the same parametric family $f(x, \Theta)$, in which the parameter $\theta \in \Theta$ is different between the groups.

Groups can be indicized by a discrete variable S to values into $1, \dots, g$. The probability to choose the group $S = j$ with $j \in (1 \dots g)$ is indicated with $l(j)$ and is equal to $\alpha_j (j = 1 \dots g)$. The joint probability density $p(x, j)$ is:

$$p(x, j) = p(x|j)l(j) = f(x|\theta_j)l(j) \quad (5.12)$$

Where in each mixture model we observe only the random variable X . The marginal density of X is:

$$p(x; \Psi) = \sum_{j=1}^g p(x, j) = \alpha_1 f(x|\theta_1) + \dots \alpha_g f(x|\theta_g) \quad (5.13)$$

Where in Ψ is a vector with all the parameters of the model. In that sense in a specific temporal interval t we can have elements belonging to different groups g , for example due to some effects. So for each time t we can consider different mixture models. These types of models can be important in financial data where mixture models seem to be very useful for returns.

5.7 Conversions between Density Data and other types of data

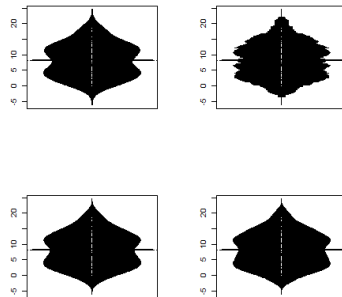
At the same time, it is possible to transform the density data into other types of data or representations like histograms (Algorithm 1.) or intervals etc. In particular following Wand Jones 1995 [687] and Sheather and Jones 1991 [624], we can obtain the binned approximation to the kernel estimate of the density functional [687]. So we can translate a specific beanplot or density data into a specific histogram.

It is important to note that in the case of the interval we are specifically losing information (and we are assuming the density is uniform). At the same time it is possible to consider the plug-in methodology to decide a bin width of the histogram and the bandwidth of the density, so it is possible to consider both techniques (histogram and density). See in this sense Scott (1979) [618], Sheather and Jones (1991) [624] and also Wand (1995) [686].

5.8 Simulation Study: effects of the kernel and the bandwidth choice

We simulate in this case various datasets to compare the different results by considering different kernels and bandwidth choices. The result of the simulation study is that there is no particular difference in choosing the various kernels in the analysis whereas it is crucial to choose a bandwidth that could be considered nearest to the optimum (figure 5.4, figure 5.5, figure 5.6, figure 5.7, figure 5.8)

Figure 5.4: Effect of the kernel and the bandwidth choice



5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

Figure 5.5: Effect of the kernel and the bandwidth choice

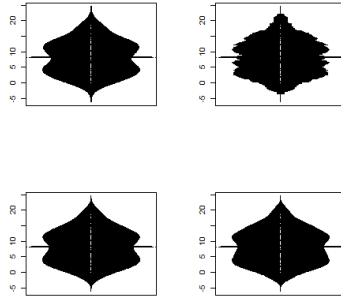
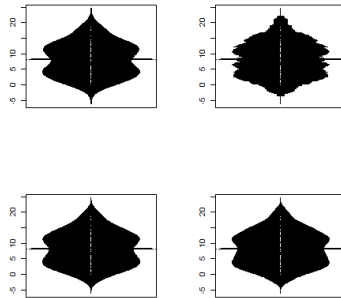


Figure 5.6: Effect of the kernel and the bandwidth choice (2)



5.9 Application on Real Data: Analysing Risk Profiles on Financial Data

We compute the histogram for the log returns of the Dow Jones Market, so we have:

Figure 5.7: Effect of the kernel and the bandwidth choice (3)

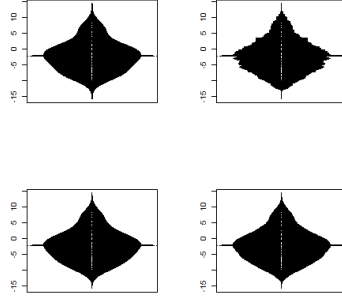
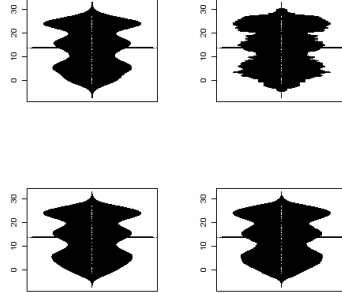


Figure 5.8: Effect of the kernel and the bandwidth choice (4)



5.9.1 Analysis of the Dow Jones Index

$$LR = \log(r) - \log(r(t-1)) \quad (5.14)$$

At this point, for the period considered we compute the kernel density estimation to observe and to compare the different subperiods. The results are coherent with the financial theory, see in this sense Di Fonso Lisi 2005 [184] and Carmona 2004 [120]

5.9.2 Analysis of the financial crisis in the US 2008-2011

The Data comes from 2 January 2007 to 12 August 2011. The data are related to the close price for each market (see the the table of symbols and markets on page 140). In particular, for each market the difference is computed.

$$DCP = CP(t) - CP(t - 1) \quad (5.15)$$

Where the differenced logarithms are computed for the most relevant countries. The density data are computed and visualized (figure 5.9, figure 5.10, figure 5.11, figure 5.12, figure 5.13, figure 5.14) whereas in the tables the quantiles for the differenced close prices are computed. As a preprocessing phase the different missing data for each market are imputed. The observations from 1 January and 13 August are considered for 2011. The analysis follows these phases:

1. Quantile Analysis for various world markets (Table 5.1)
2. Analysis of the Density Data 2007-2011 and the Profile Risk Indicator (The 5% quantile of the difference between two consecutive values or the daily variation by year) for various world markets
3. Comparison between the Density Bandwidth (in figure 5.16) and an Index of Market Uncertainty as the VIX Index 1990-2011 in figure 5.17 (see in this sense Bloom 2009 [89])
4. Comparing the result with the Radius of the Interval Data 1990-2011 (figure 5.15)

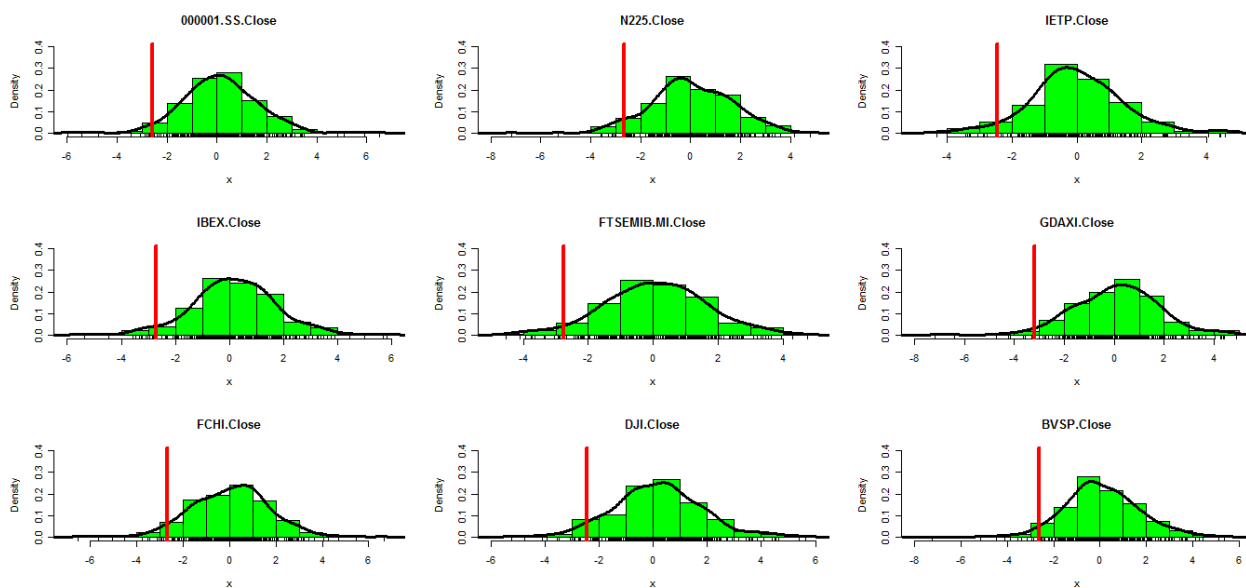
In practice we can observe the different evolution of the Dow Jones market related to other markets. The differenced value presents some

interesting features related to the volatility that could be effectively captured by the density data. For each year we can note that there are some economic situations (US, Italy, France, Germany, Brazil for example) that in different situations are behaving differently and can have different market responses to the shocks (see Bloom 2009 [89]). In particular it seems possible to observe, considering the bandwidth sequence extracted year by year from the US data, that they are capable of identifying well the most relevant financial shocks on the market (see in this sense [90]). It is at the same time confirmed that the methods considered here allow one to observe phenomena like the implied volatility over the time and perhaps anticipate them by using some forecasting methods over time. The use of density permits one to observe the entire data structure, as will be seen in the following chapters.

5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

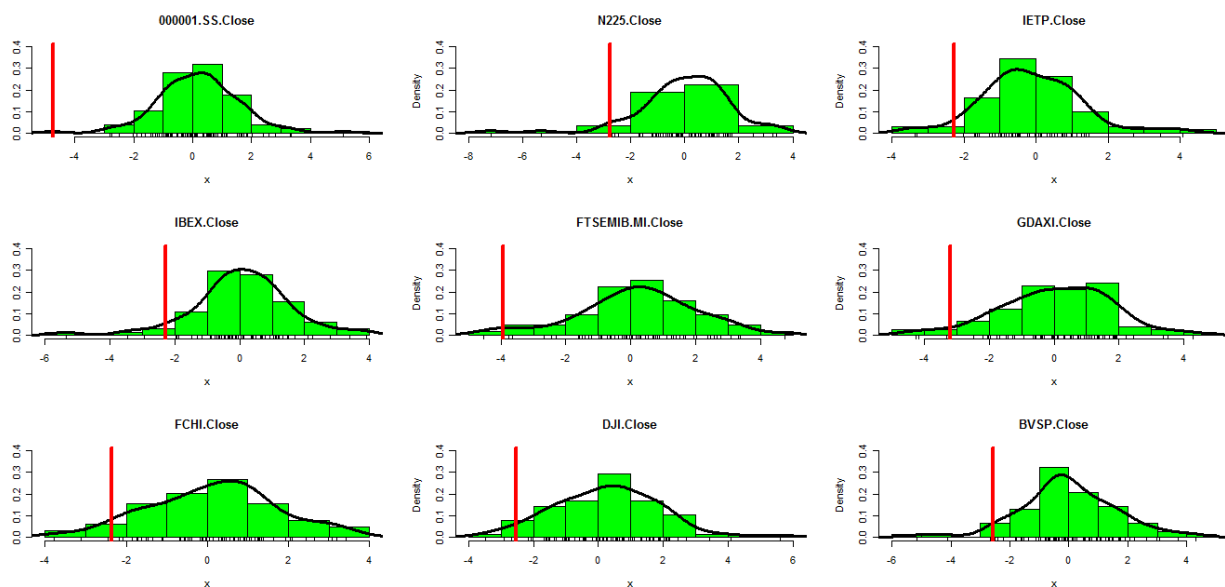
Summary Results: Densities (Kernel Density Estimation)
An Internal Representation preserving complex patterns of intra-data variation (like Histograms)
More flexible than Histograms
More interpretable than Histograms
Preserve continuity of the data (without representing it bin by bin).
Need to carefully decide the bandwidth.

Figure 5.9: Density Estimation and Profile Risk Indicator computed



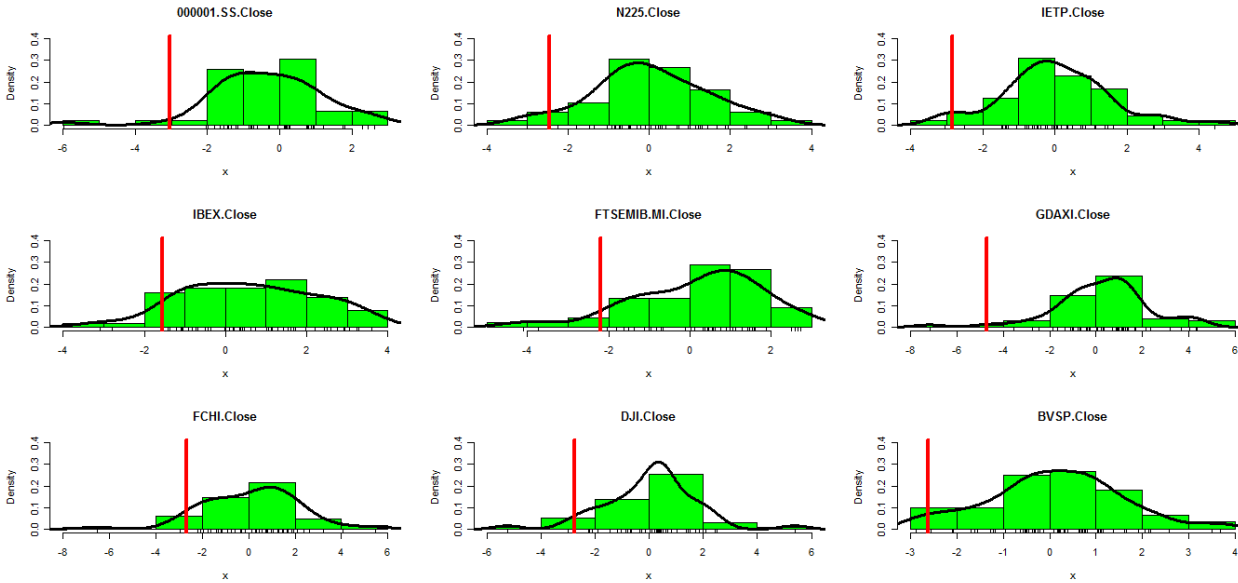
FOUNDATIONS OF DENSITY VALUED DATA: REPRESENTATIONS

Figure 5.10: Density Estimation and Profile Risk Indicator year: 2007



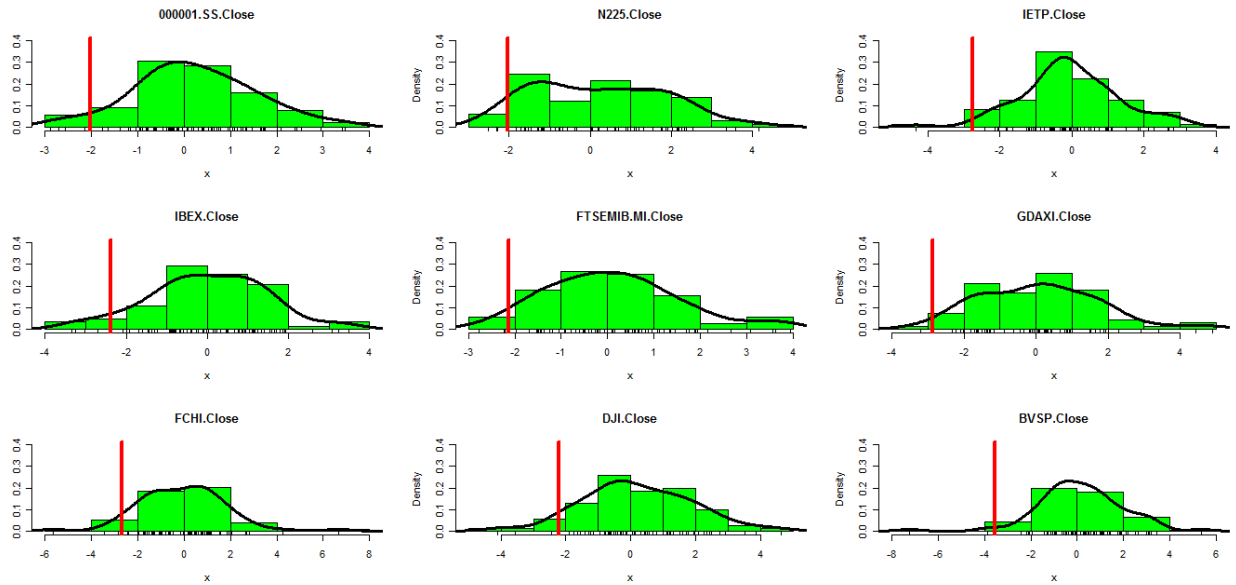
5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

Figure 5.11: Density Estimation and Profile Risk Indicator computed year: 2008



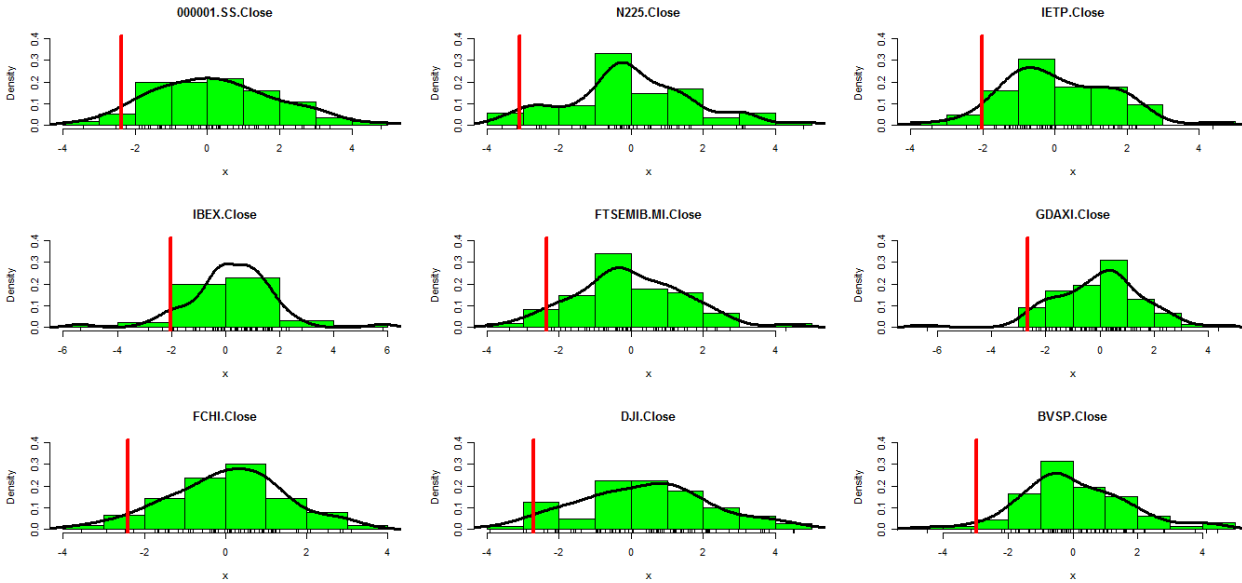
FOUNDATIONS OF DENSITY VALUED DATA: REPRESENTATIONS

Figure 5.12: Density Estimation and Profile Risk Indicator year: 2009



5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

Figure 5.13: Density Estimation and Profile Risk Indicator computed year: 2010



FOUNDATIONS OF DENSITY VALUED DATA: REPRESENTATIONS

Figure 5.14: Density Estimation and Profile Risk Indicator computed year: 2011

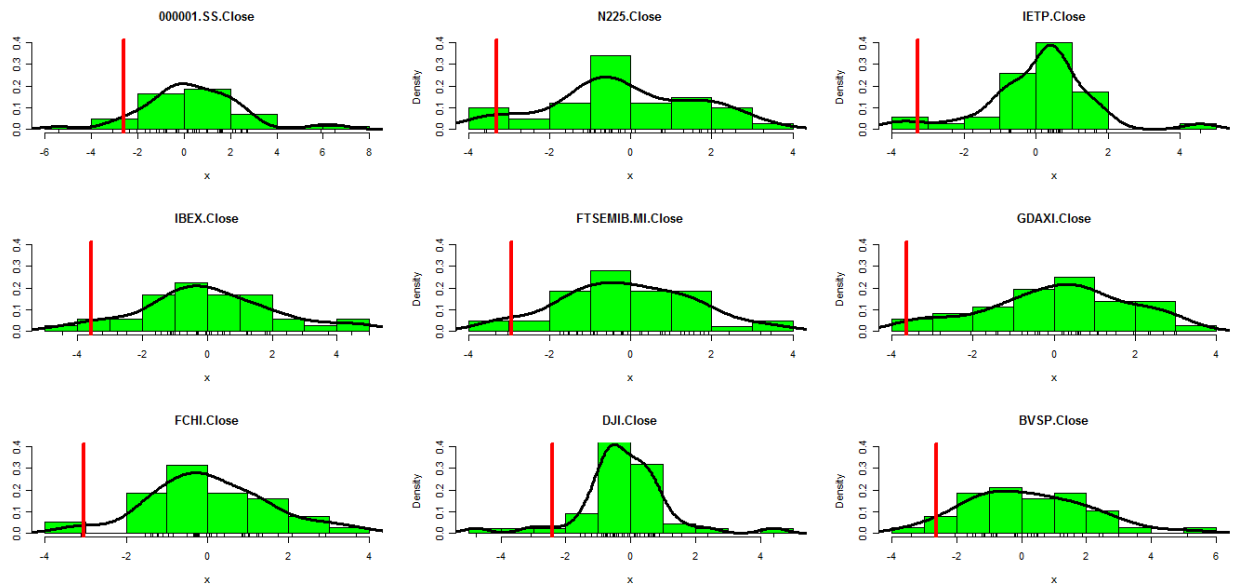
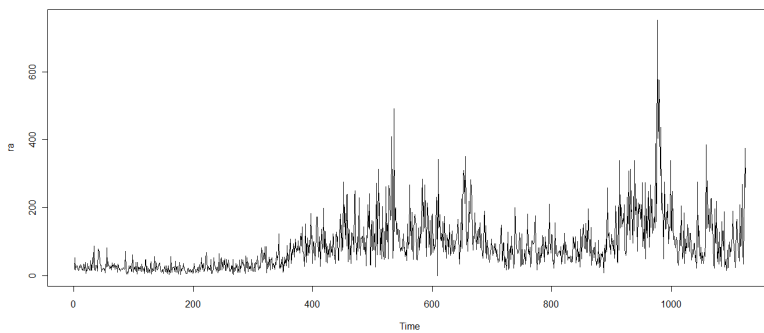


Figure 5.15: Radius of the Interval time series (ITS) DJI 1990-2011



5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

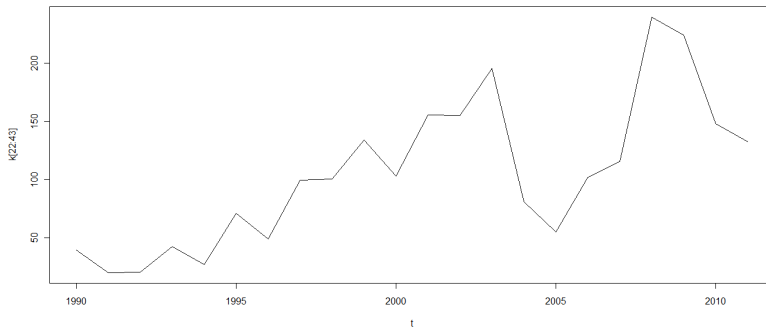
Table 5.1: Risk profiles: quantiles computed 2007-2009

	0%	25%	50%	75%	100%
TA100	-81.13	-8.27	0.03	8.83	57.57
EGX30.CA	-1162.36	-47.40	0.00	60.66	2140.74
GSPTSE	-864.41	-73.20	12.09	93.03	890.50
MXX	-1957.74	-202.18	29.02	226.13	2190.62
MERV	-329.20	-16.56	2.34	22.14	154.74
STI	-294.44	-19.79	0.35	22.06	191.67
KS11	-126.50	-11.66	1.98	14.70	115.75
NSEI	-496.50	-41.97	2.55	46.33	651.50
NZ50	-139.95	-14.92	0.61	14.86	166.58
KLSE	-227.66	-4.85	0.88	6.31	235.67
JKSE	-200.45	-16.40	4.06	22.91	182.66
BSESN	-1408.35	-145.24	11.80	155.10	2110.79
AORD	-408.90	-34.78	1.80	37.37	280.50
RUA	-60.26	-4.73	0.63	5.13	59.78
RUT	-63.67	-6.88	0.78	6.94	48.41
RUI	-57.35	-4.48	0.58	4.65	57.81
SPSUPX	-23.66	-2.22	0.18	2.17	23.30
SML	-33.06	-3.45	0.37	3.54	23.86
MID	-69.71	-6.28	1.02	7.04	57.60
GSPC	-106.85	-8.05	1.08	8.33	104.13
NDX	-175.89	-12.84	2.15	15.36	159.74
IXIC	-199.61	-16.59	2.65	19.26	194.74
NIN	-270.21	-39.49	8.38	58.42	266.53
NYA	-686.36	-54.60	6.85	60.54	696.83
DJU	-31.14	-2.47	0.36	3.00	46.01
DJT	-399.19	-45.01	4.20	45.76	298.05
DJA	-274.42	-25.13	2.88	27.22	319.16
FTSE	-391.10	-40.20	1.00	42.80	431.30
SSMI	-451.60	-46.65	0.00	47.70	609.10
OMXSPI	-17.63	-2.61	0.07	2.60	20.42
OSEAX	-32.14	-4.12	0.50	4.43	33.13
OMXC20.CO	-38.88	-3.40	0.13	3.08	27.96

FOUNDATIONS OF DENSITY VALUED DATA: REPRESENTATIONS

	0%	25%	50%	75%	100%
BFX	-224.64	-24.38	0.06	21.31	268.92
ATX	-241.38	-32.65	0.17	30.85	331.51
AEX	-31.46	-2.77	-0.02	2.83	30.17
X000001.SS	-354.69	-27.10	3.27	36.00	351.40
N225	-1089.02	-105.53	3.60	107.22	1171.14
IETP	-79.19	-6.69	-0.28	5.94	64.25
IBEX	-1029.60	-106.30	2.50	93.80	1305.80
FTSEMIB.MI	-2135.00	-244.49	2.00	223.00	2333.00
GDAXI	-523.98	-46.90	3.53	48.01	518.14
FCHI	-368.77	-35.95	-0.35	34.80	367.01
DJI	-777.68	-62.40	6.26	69.33	936.42
BVSP	-4755.00	-519.00	77.00	605.00	5219.00

Figure 5.16: Bandwidth for the US Densities computed over the years



5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

Table 5.2: International Stockmarket Symbols

	Symbol	Country
1	TA100	Israel
2	EGX30.CA	Egypt
3	GSPTSE	Canada
4	MXX	Mexico
5	MERV	Argentina
6	STI	Singapore
7	KS11	South Korea
8	NSEI	India
9	NZ50	New Zealand
10	KLSE	Malaysia
11	JKSE	Thailand
12	BSESN	India
13	AORD	Australia
14	RUA	USA
15	RUT	USA
16	RUI	USA
17	SPSUPX	USA
18	SML	USA
19	MID	USA
20	GSPC	USA
21	NDX	USA
22	IXIC	USA
23	NIN	USA
24	NYA	USA
25	DJU	USA
26	DJT	USA
27	DJA	USA

	Symbol	Country
28	FTSE	United Kingdom
29	SSMI	Switzerland
30	OMXSPI	Stockholm
31	OSEAX	Norway
32	OMXC20.CO	Denmark
33	BFX	Belgium
34	ATX	Austria
35	AEX	Netherlands
36	X000001.SS	China
37	N225	Japan
38	IETP	Ireland
39	IBEX	Spain
40	FTSEMIB.MI	Italy
41	GDAXI	Germany
42	FCHI	France
43	DJI	USA
44	BVSP	Brazil

5.9. Application on Real Data: Analysing Risk Profiles on Financial Data

Figure 5.17: Implied Volatility for the US Market (VIX Index Index of volatility expectations (Bloom 2009 [89])

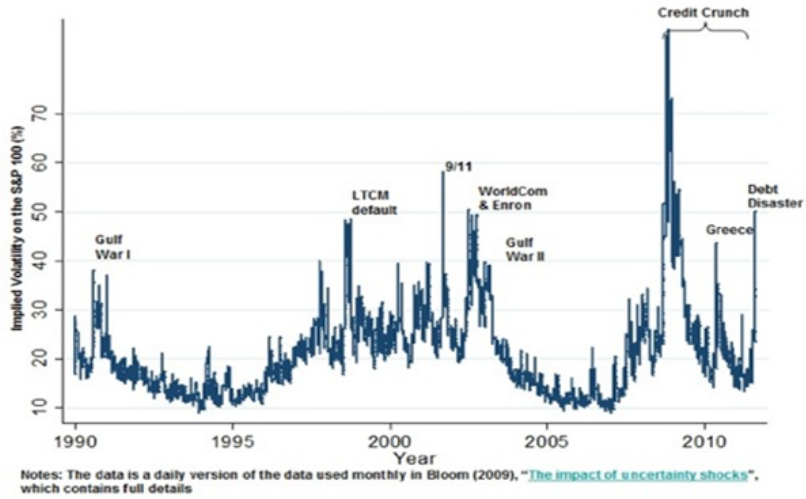
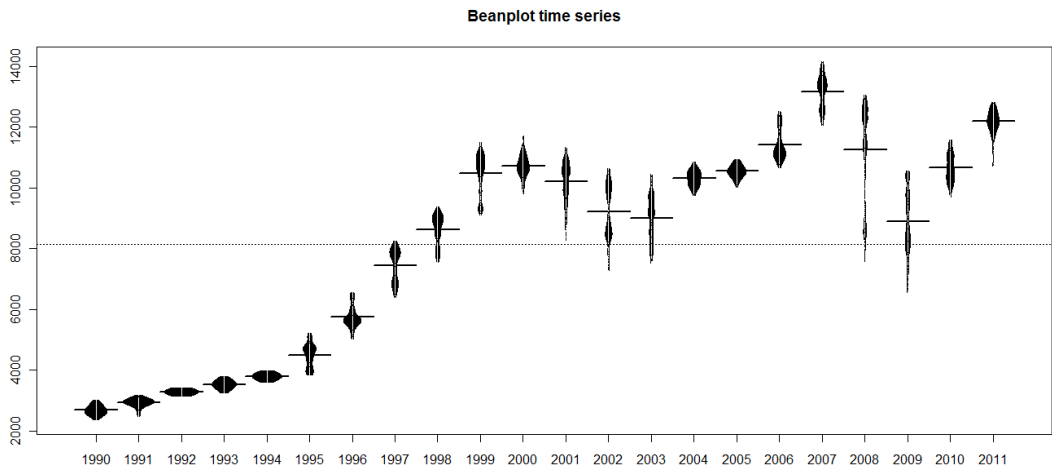


Figure 5.18: Beanplot Time series (BTS) DJI 1990-2001



Part II

New Developments and New Methods

Chapter 6

Visualization and Exploratory Analysis of Beanplot Data

In this chapter we propose a new approach for the aggregation, the visualization and the analysis of the complex time series¹ seen in Chapter 2. In particular, this approach is based entirely on a representation like the density data² or the beanplot data (Kampstra (2008) [416]).

Thus we are in the framework of Chapters 3, 4 and 5 where we tried to summarize our data by considering classical intra-period statistical representations. These types of new aggregated time series (Beanplot time series BTS) can be successfully used when there is an overwhelm-

¹By now we adopt the definition of complex time series in Diday 2002 [214] who defines the complex time series and the adequate description of the subperiod: "representing each time series by the histogram of its values or in describing intervals of time". In this respect we consider the intra-period representations as genuine representations of the phenomena

²Clearly the density data comes from the kernel density estimation seen in Chapter 5

ing number of observations, for example in High Frequency financial data (in particular we refer to the specific characteristics of these data as presented in Chapter 2, see also Dacorogna et al. 2001 [163]). In these cases, it is not possible to visualize the data adequately and so it is necessary to maximize the information obtained by considering another intra-period representation³.

A second reason to use these types of aggregate representations is that they allow the "uncertainty visualization" (Griethe Schumann 2006 [326] Johnson 2004 [408] and Potter 2006 [564]) where in the original data there are present: "error, imprecision, lineage, subjectivity, noise, etc".⁴ It is possible to visualize the uncertainty adequately by considering some alternative representations which have been proposed in literature in recent years.

As we showed in Chapter 2, there are important cases in complex data (for example, the financial time series) in which there is uncertainty, due to the structure of the data for example⁵ (errors, missing value, etc.)

A third reason can be considered to be the capacity of aggregate representations to detect patterns in data. So, they can be useful for analyzing the complex behaviour of the markets where we can discover important patterns in the long time⁶. For example, these methods can capture complex patterns of dependency over the time, where it is known that financial markets show the example phenomena of long time autocorrelation (see Cont [153], Muchnik, Bunde, & Havlin (2009) [519] and Henry Zaffaroni 2003 [358]). A similar result is obtained using the Histogram Data of González-Rivera and Arroyo 2011

³See Tufte 1983 [668]

⁴Boller Braun Miles and Laidlaw 2010 [101]; for a different and opposing opinion about the uncertainty visualization see Boukhelifa Duke 2009 [105]

⁵See Brownlees and Gallo 2006 [115] but also Dacorogna et al. 2001 [161]

⁶In particular, one of the reasons to use these types of representations is the capability to analyse the long run dynamics of the complex time series

[318] which find some correlations between low values of the market index over time. In general the Econophysics approach analyses the complex behavior of the markets in the data⁷.

These types of complex dependencies on data could be analysed by considering the minima and the maxima of the beanplot time series BTS⁸. The beanplot time series BTS naturally represents the internal variation and the uncertainty, whereas the stripchart (the internal observation) reduces the information.⁹ It is important to note that the use of the beanplot enhances the possibility to compare a higher quantity of data, and so could be useful in the analysis of the risk over time¹⁰

We present in this chapter a representation based on the beanplot time series BTS, of high frequency data, while in the next we propose a particular coefficient estimation (see Chapter 7) and we will use it for the forecasting and clustering aims later in the work (Chapter 8 and 9). In particular, we will show the usefulness of this approach in analysing the long run dynamics of the markets and the business cycle.

This chapter is organized as follows: in the first part we approach the economic problem and how it is possible to obtain density data types of data by starting from a different type of scalar data. In the third part we present the beanplot or density data and we describe

⁷In recent years there has been the growth of many works that try to understand the financial markets as complex systems, see Mantegna and Stanley 2000 a first approach into Econophysics [485] and for a view of the Financial Markets as real world complex system see Johnson Jeffries Ming Hui 2003 [409]. On the criticism of the main classical economic models and for the search of a new paradigm see Mandelbrot Hudson 2006 [483]

⁸For example, it is interesting to analyse the complex characteristics in a financial time series: see Cont 2001 [152] and Sewell 2008 [619]

⁹There are important reasons to choose the interval data in the analysis

¹⁰For example in the analysis of Risk and Financial Risk it could be very important (see Resti Sironi 2007 [580])

their features. In the fourth part we introduce the beanplot time series (BTS) and the possibility to gain new information with these types of data from the characteristics of the scalar time series. In the fifth and the sixth parts we introduce the internal and the external modelling process, whilst in the Ninth Chapter we introduce the mixture analysis as a tool in the diagnostic internal modelling.

6.1 The Data Aggregation problem

In particular an (ordinal) scalar time series takes the form $\{y_t\}, t = 1...T$ with $y_t \in \mathfrak{R}$ and can have a single value in \mathfrak{R} .

By considering more than one statistical unit $\{y_{t,i}\}, t = 1...T, i = 1...I$ we can consider longitudinal data (time series cross sectional data)¹¹. Usually cross sectional time series are different from the panel data because here I is fixed and T tends to be large whereas in the panel data case I is large and T is fixed.

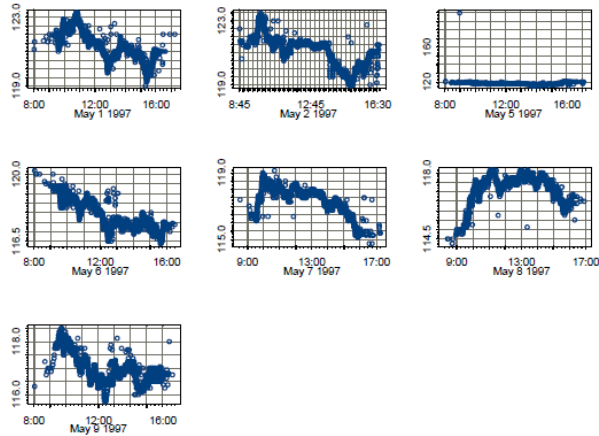
There are real cases, in particular, in which scalar time series y_t does not allow one to correctly approach a phenomena, in particular when the dataset contains a huge quantity of observations and their visualization is not possible (see figure 6.1). Another important case could happen when we are interested not in a single value but in a specific distribution of a variable K in a given temporal interval T (for example, Arroyo and Matè (2006) [43] refer to the variable as outcomes of the daily time-varying demand of energy). In other works, Arroyo et al. (2011) [38] try to predict the histogram data over time¹².

¹¹Beck 2004 [69] shows a high quantity of examples of longitudinal data, all these examples are characterised by a higher quantity of information than "normal" datasets

¹²In all these approaches, authors directly consider the Symbolic Data Analysis approach as considered in Diday 2002 [206]

6.1. The Data Aggregation problem

Figure 6.1: Intra-day price data for Microsoft stock (Zivot 2005 [722])



In all these cases we are trying to forecast distributions at T whereas in a different (in scalar time series analysis) way we would force them to be a single value, for example by aggregating the values. The case is typical in high frequency financial datasets in which data are collected at a given high frequency (for example, minutes), but sometimes they need to be analyzed at a lower frequency (daily): in this case it is necessary to aggregate the data using a statistical method and minimize the information loss due to the aggregation¹³. In other cases the data

¹³Goodfriend 1992 [316] shows that "aggregation in the presence of data processing lags distorts the information related to the data". In this case the aggregation became really problematic. At the same time, Dacorogna et. al. 2001 reports that empirical results in financial data can change, by considering different types of data [163] (pag.143). The problem of the effects of the aggregation data and the information loss associated with it is highly debated in literature starting from Orcutt Watts Edwards 1968 [540]. An interesting conclusion is also in McKee and Miljkovic 2007 [498] "aggregate series are appropriate for long-term decision analysis, but some information loss occurs when conducting short-term decision analysis"

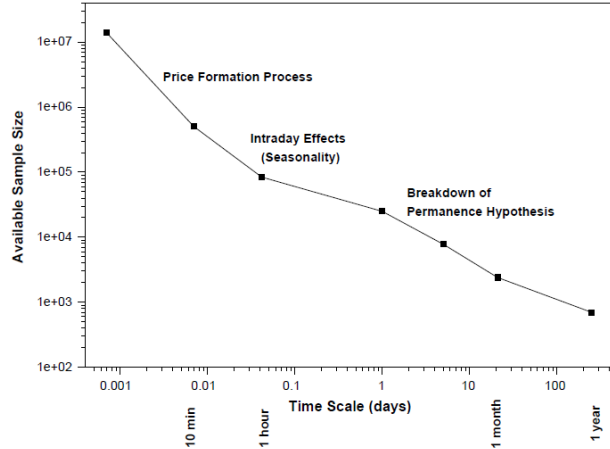
are aggregated by considering simply the last value (see Dacorogna et al. 2001 [163]). The aggregation process does not faithfully represent the intra-day dynamics where data are observed only at some equilibrium levels and it is neglected when these equilibrium values are reached (Engle and Russel (2004)[253]). In fact the aggregation does not represent correctly the underlying phenomenon and a time series of distributions can be more useful than the other forms of aggregated time series (Arroyo, Gonzales Rivera and Matè (2009) [40]).

The problem is increasingly important considering the growing size of the modern datasets. In particular Schweitzer says that: “Distributions are the number of the future!” (Schweizer (1984) [615]), so there is the possibility, followed in literature, to cope directly with the distributions but not with the original data.

Various approaches in this sense were followed in literature to obtain appropriate data representations. Different methods can be either parametric or nonparametric one. Arroyo, Gonzales Rivera and Matè (2009) [40], propose Histograms as nonparametric method. This approach can be related directly to the Symbolic Data Analysis (for other approaches considering complex time series see Diday and Noirhomme (2008) [218] Billard and Diday (2003) [86]). In this respect Symbolic Data Analysis proposes an alternative way to manage huge datasets: transforming the original data into symbolic data as Intervals, Histograms, Lists, etc., by retaining the key knowledge¹⁴. In these symbolic datasets, items are described by symbolic variables (Arroyo Gonzales and Rivera and Maté (2009) [40]) and the cells can contain entire distributions (Diday (2006))

¹⁴See for example Billard and Diday 2010 [88]

Figure 6.2: Financial data types with the typical sizes and frequency (Dacorogna et al. 2001) [163]



6.1.1 High Frequency Data and Intra-Period Variability

Here we deal in depth with the problem of visualizing and exploring specific beanplot time series (BTS) deriving from high-frequency financial data (see in figure 6.2 and in figure 6.3 for its characteristics as frequency and its typical use in Finance and Economics). These data present unique features, absent in low frequency time series, which involve the necessity of searching and analysing an aggregate behaviour. Infact these data are typically overwhelming and they tend to neglect for example the data visualization (see also Drago, Scepi 2009 [236]). Therefore, we obtain from the original data a particular aggregated time series called a beanplot time series (BTS). We show the advantages of using these instead of scalar time series when the data shows a complex structure. Furthermore, we underline the interpretative proprieties of beanplot time series (BTS) by comparing different types of

aggregated time series. In particular, with simulated and real examples, we will illustrate the different statistical performances of beanplot time series (BTS) in respect to boxplot time series (BoTS).

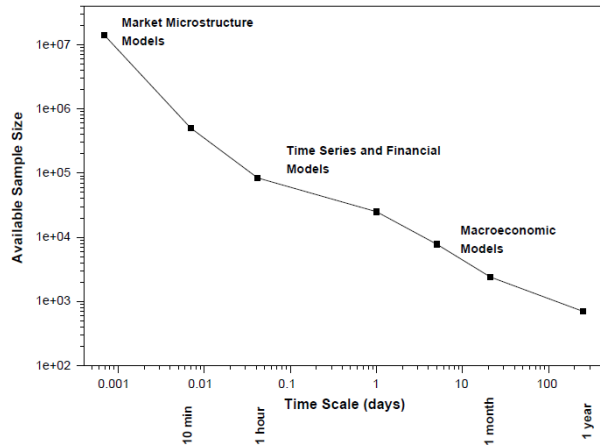
High-frequency financial data (Engle Russell 1998 [253]) are observations on financial variables collected daily or at a finer time scale (such as time stamped transaction-by-transaction, tick-by-tick data, etc.). This type of data have been widely used to study various market microstructure related issues, including price discovery, competition among related markets, strategic behaviour of market participants, and modelling of real-time market dynamics. Moreover, high-frequency data are also useful for studying the statistical properties, volatility in particular, of asset returns at lower frequencies. The analysis of these data is complicated for different reasons. We deal with a huge number of observations ("the average daily number of quotes in the USD/EUR spot market could easily exceed 20,000" see Engle Russell 2009 [255]), often spaced irregularly over time, with diurnal patterns, price discreteness, and with a complex structure of dependence. The characteristics of these data do not allow the visualization and exploration by the means of classical scalar time series. Furthermore, it becomes very difficult to forecast data without defining an aggregate behaviour.

In this chapter we will introduce beanplot time series (BTS) with the aim of synthesizing and visualizing high-frequency financial data or, more in general, complex types of temporal data. In particular, we will discuss their properties by proposing critical comparisons among different possible aggregated time series. After that, we will carry out several simulated examples (in section 6.7), starting from different models, different numbers of observations and different intervals of aggregation to show how beanplot time series (BTS) tend to perform better than boxplot time series (BoTS). Some interpretative rules are given in subsection 6.7.1. We have enriched our analysis by an application on real high frequency financial data where we show how

6.1. The Data Aggregation problem

beanplot time series (BTS) easily detect real intra-day patterns (in section 6.9).

Figure 6.3: Financial data models with the data typical sizes and frequency (Dacorogna et al. 2001) [163]



6.1.2 Representations, Aggregation and Information Loss

It is a known fact that time series databases in various fields are very large (see for example Lin Keogh and Lonardi 2007 [422] Nguyen Duong 2007 [537]). In that sense, the data extraction could be very difficult and visualization inefficient. So it is necessary to represent the data in a way that adequately manages these problems. In particular, we consider that in literature there exist two types of transformations based on dimensionality¹⁵ and numerosity reduction techniques¹⁶. In

¹⁵See Gunopulos 2011 [329] and Lin Keogh and Lonardi 2007 [422] and Wang Megalooikonomou 2008 [768]

¹⁶See Pekalska, Duin et al. 2006 [554] Wilson and Martinez 1997 [772]

this sense, the literature is very rich and symbolic representations¹⁷ are in this sense linked to two different approaches (Lin Keogh and Lonardi 2007 [422]):

1. Data Adaptive

- (a) Sorted Coefficients
- (b) Piecewise Polynomial
- (c) Singular Value Decomposition
- (d) Symbolic
- (e) Trees

2. Non Data Adaptive

- (a) Wavelets
- (b) Random Mappings
- (c) Spectral
- (d) Piecewise Aggregate Approximation

In that sense, as we have already seen, there is another approach which considers the problem of finding an adequate internal representation of the variability.

In particular the same data, for each observation, can be considered as characterized by internal variability as well (complex time series and high frequency data for example). In that case, data are not scalars

¹⁷Lin Keogh Lonardi Chiu (2003) [459]

but are Intervals, Histograms, Boxplots etc.

So, there are also cases where these internal representations are interesting on their own (for example where there is the interest in modelling the intra-period variability). In those cases we want to explicitly analyse this variability.

More in general with the aim of summarizing and visualizing high frequency time series, different types of statistical tools and aggregated time series can be considered¹⁸.

In particular, we can consider with respect to the problem of data aggregation in complex time series these solutions:

1. Intervals
2. Boxplots
3. Histograms
4. Candlesticks
5. Beanplots

We define these as Complex Objects where in some cases they can be defined as Temporal Symbolic Data (usually Symbolic Representations are defined as the representation of time series using some other types of the methods seen above).

In practice we are trying to represent and to visualize the statistical uncertainty by considering different Complex Objects (see Potter 2006 [564]). In massive data sets the Data Analysis process can be conducted in two different ways: considering the original data using the classical temporal data mining techniques¹⁹ or the aggregate representations (or the Complex Objects).

¹⁸In all these cases we can use stripcharts, boxplots, histograms and so on

¹⁹See for some overviews on the topic Laxman, Srivatsan, and P S Sastry. 2006 [441], Antunes and Oliveira 2001 [23] and Mitsa 2010 [510]

More generally identifying the data structure can be very relevant to understand the phenomenon from a statistical point of view²⁰. The mean, the median, or the total of the single values represent weak aggregations because important information is neglected (see Marvasti 2011 [489]).

Initially, we used a stripchart time series (that could be related to the Interval Time Series (ITS): see Arroyo and Maté 2006 [43]). This type of time series correctly shows the original trend as well as the minimum and the maximum of each interval (a day, for example). However in such graphics, one dot is plotted for each observation in the single time interval and, consequently, it is a useful tool only when there are very few points.

Therefore, it might be difficult to apply them in the high frequency data framework. A recent proposal (see Arroyo and Maté 2009 [44] and Arroyo et al.[37]), in the context of symbolic data, consists of substituting time series of observations with histogram time series HTS (see Arroyo et al. 2011 [38]). These representations are very useful for temporal and spatial aggregations for many reasons: they are simple with flexible structure, and they have the capacity to describe the essential features of the data with reasonable accuracy and with closeness to the data, without imposing any distribution. Nevertheless, the multiple histograms are difficult to compare when there are many of them plotted on a graph, because the space becomes cluttered²¹.

Tukey's boxplot (see Tukey 1977 [670]) is commonly used for com-

²⁰Jackson 2008 [401] wrote "The casual use of hypothesis tests based on arbitrary thresholds is frequently criticized (Gelman and Stern 2006 [294]), particularly in medical research (Sterne et al. 2001 [642])"

²¹Another difficulty in using histograms is given by Elgammal et al. 2002 [246]: "The major drawback with color histograms is the lack of convergence to the right density function if the data set is small... Unlike histograms, even with a small number of samples, kernel density estimation leads to a smooth, continuous, and differentiable density estimate".

paring distributions between groups. For time series data, the boxplot seems to show several features of the temporal aggregation: center, spread, asymmetry and outliers. Furthermore, Box Plot time series detect main structural changes well (Maté Arroyo 2006 [493]). However the number of outliers detected will increase if the number of observations grows and the information about the density is neglected. This information can be very important in the aggregation of high frequency financial data where different volatility clusters can arise.

A way to visualize the data uncertainty related to the presence of different bumps by enhancing the initial boxplots is given by Hyndman with the HDR boxplots (Hyndman 1996 [378]). Using these tools we are able to identify the regions of highest density and so bumps can occur in the density as well.

In order to retain this information, it is possible to use at the same time the Violin Plot (see fig. 6.4) time series. This tool (Benjamini 1988 [73]) combines the advantages of boxplots with the visualization of the density and it provides a better indication of the shape of the distribution. However, in a Violin Plot²² the underlying distribution is visible but the individual points, besides the minimum and maximum, are not visible and no indication of the number of observations in each group is given.

Other proposals related to the uncertainty in representations are contained in Jackson (2008) [401] and Cleveland (1993) [146]. In particular the first author makes a proposal related to the density strip, in which data are represented as a thin horizontal rectangle which is darkest at the point of highest probability density, white at points of zero density, and shaded with darkness proportional to the density²³. In this case the relevant dimension is given by the shade of the image.

²²There was in literature another proposal in this sense from Messing 2010 [815] that considers a single weight for each observation

²³Jackson 2008 [401]

At the same time in violin plots there is not a specific visualization of the single observations (so it is not possible to clearly identify the outliers in data). Another proposal in literature is the Summary Plot by Potter Kniss Riesenfeld 2007 [565] in which we observe the density, the moments and the cumulant information

Other alternatives that use shading in visualizations to show densities are in Cohen Cohen 2006 [148] using the Sectioned density plots.

Our proposal consists of using Beanplot time series -BTS (Kampstra 2008 [416]) in particular in the context of high frequency financial data. Indeed, in each single beanplot all the individual observations are visible as small lines in a one-dimensional scatter plot, as in a stripchart²⁴. In the Beanplot time series (BTS), both the average for each time interval (represented by the beanline) and the overall average is drawn; this allows for an easy comparison among temporal aggregations. The estimated density of the distribution is visible and this shows the existence of clusters in the data and highlights the peaks, valleys and bumps. Furthermore, anomalies in the data, such as bimodal distributions, are easily identified.

This is very interesting information in the context of high frequency financial time series where the intra-period variability represents the main characteristics of the data. The number of bumps can be considered as a signal of different market phases in the daily market structure. We can also observe that the beanplot becomes longer in the presence of price anomalies such as peculiar market behaviours (speculative bubbles). See figures 6.4 and 6.5 for the different comparative object that could be used to represent the intra-period dynamics of scalar complex time series. At the same time, in the following table we summarize all the different complex objects that could be considered

²⁴ Eklund 2010 [780] proposes a different tool like the beeswarm that can be considered an improved stripchart where it is possible to visualize all the observation (experimentally on a limit of 1000-2000 observations)

in literature to be useful to represent intra-period variations in temporal data analyses (the complex objects). In the following table, there is also a review of the complex objects that could be considered in literature and their linkages with Symbolic Data Analysis as symbolic data²⁵.

Complex Object	Reference	Data
Stripchart	Dalgaard 2002 [166]	Interval
Beewarms	Eklund 2010 [780]	Interval
Standard Boxplot	Tukey 1977 [670]	Boxplot
HDR Boxplot	Hyndman 2006 [378]	Boxplot
Box Percentile Plot	Esty Banfield 2003 [257]	Boxplot
Histogram	Pearson 1895 [553]	Histogram
Summary Plot	Potter Kniss Riesenfeld 2007 [565]	Density
Sectioned Density Plot	Cohen Cohen 2006 [148]	Density
Violin Plot	Adler 2005 [6]	Density
Weighted Violin Plot	Messing 2010 [815]	Density
Beanplot	Kampstra 2008 [779]	Beanplot
The different complex objects in literature		

6.2 From Scalar Data to Beanplot Data

We start from a classical time series $\{y_t\}, t = 1 \dots T$ with $y_t \in \mathfrak{R}$ and a single value in \mathfrak{R} . This time series can generate a symbolic one (or a time series of aggregate representations) by contemporaneous or temporal aggregation from the original scalar time series y_t part of a set S .

In the contemporaneous aggregation we have a sample of n time series denoted as y_{it} , where i is related to a different statistical unit,

²⁵Piccolo 2000 offers an interesting review of the graphical exploratory methods in Statistics [562]

Figure 6.4: The evolution from the boxplot (Harrell et al. 2011 [340])

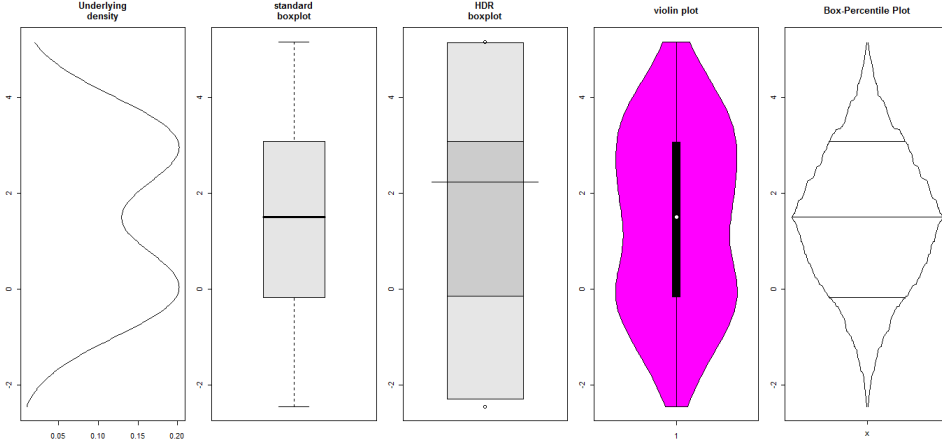
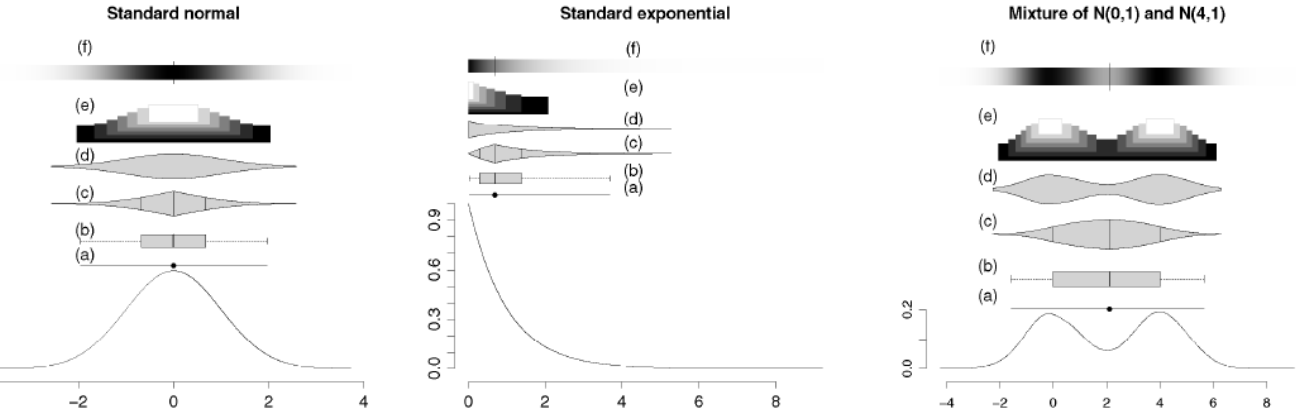


Figure 6.5: Uncertainty in Representations



so we aggregate the data either by units i , or time t .

In the temporal aggregation (see Arroyo 2009 [32]) we aggregate a specific time series y_t part of S only by considering the time t . For

example, we can obtain as a symbolic time series a portfolio of stocks relying on a specific criteria or sector (contemporaneous aggregation) or a financial high frequency time series considered at a lower frequency (temporal aggregation).

Definition 1. A primary data is an element x_i in the set S . A secondary data is a representation of the set S and they can be represented as complex data (Billard and Diday 2006 [86]).

The statistical analysis can be conducted in two parallel ways: on primary data (classical time series) and on the secondary data (the time series representing portfolios, for example).

6.3 Beanplot Data

Following Arroyo, González-Rivera and Maté (2006), by taking into account a variable X (for example the closing price of a stock) we consider as primary, single observations part of a set S . For every element $x_i \in S$ we observe a secondary datum as a density.

Definition 2. A density data at time t is a representation of the x_i single elements in the set S , such as from Chapter 5:

$$\hat{f}_{h,t} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (6.1)$$

where K is a Kernel and a h is a smoothing parameter defined as a bandwidth. K can be a gaussian function with mean zero and variance 1. The Kernel as we know is a non-negative and real-valued function $K(z)$ satisfying: $\int K(z)dz = 1$, $\int zK(z)dz = 0$, $\int z^2K(z) = k_2 < \infty$ with the lower and upper limits of integration being $-\infty$ and $+\infty$. It is possible to use various Kernel functions (Ke): uniform, triangle, epanechnikov, quartic (biweight), tricube (triweight), gaussian and cosine. The variance can be controlled through the parameter h :

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x - x_i^2}{2h^2}} \quad (6.2)$$

Various methods have been proposed in literature to choose the bandwidth h . Jones, Marron and Sheather (1996) proposed for example a bandwidth choice, based on the standard deviation:

$$h_{SNR} = 1.06Sn^{-1/5} \quad (6.3)$$

In the data visualization we use the Sheather-Jones criteria that defines the optimal h in a data-driven choice (Kampstra 2008).

Definition 3. A Beanplot data $\{b_{Y_t}\}$ is a combination of a 1-d scatterplot²⁶ and a density trace (Kampstra 2008 [416]).

The beanplot can be considered as a particular case of interval-valued modal variable at the same time as boxplots and histograms (see Arroyo and Maté (2006)). In a beanplot we take into account both the interval between the minimum a_{L_t} , the maximum a_{U_t} and the density as the kernel nonparametric estimator (the density trace, see Kampstra (2008)). Every single observation x_{it} is represented on the 1-dimensional scatterplot. This feature is useful to visually detect observations distant from the others in the set S . The beanline at time t is a central measure of the beanplot (and a measure of location of the object) and is defined as:

$$a_{M_t} = \frac{\sum_{i=1}^n (x_{it})}{n} \quad (t = 1 \dots T) \quad (6.4)$$

An alternative centre measure a_{M_t} is the median, where the quantiles can be considered in measuring the beanplots (in particular the size of the object).

In the beanplot, the variability or size is mainly represented by the

²⁶Dalgaard 2002 [166]

6.3. Beanplot Data

interval related to the minima a_{L_t} and the maxima a_{U_t} of X . The minima and the maxima localize as well the turning points in the original time series (in figure 6.6). The beanplot size can be represented in its data interval by $[a_t]$ over the base set (E, \leq) is the ordered pair $[a_t] = [a_{L_t}, a_{U_t}]$ where $a_{L_t}, a_{U_t} \in E$ are the interval bounds such as $a_{L_t} \leq a_{U_t}$. Inside the interval $[a_{L_t}, a_{U_t}]$ represent the single primary observations (represented as a 1-dimensional scatterplot or stripchart) so we are able to understand the location of the single observations in the set S . The measure of size in the beanplot $\{b_{Y_t}\}$ is:

$$a_{S_t} = a_{U_t} - a_{L_t} \quad (t = 1 \dots T) \quad (6.5)$$

Where a_{U_t} is the upper bound and a_{L_t} is the lower bound.

At the same time, it is possible to consider the interval composed by the two consecutive sub-intervals (or half-point) through the beanline (the radii of the beanplot $\{b_{Y_t}\}$): $[a] = \langle a_{C_t}, a_{R_t} \rangle$ with:

$$a_{C_t} = \frac{(a_{U_t} + a_{L_t})}{2}, a_{R_t} = \frac{(a_{U_t} - a_{L_t})}{2} \quad (t = 1 \dots T) \quad (6.6)$$

The interval arithmetic (Moore 1966 [513]) can be applied to beanplot data.

So, why do we deal with beanplots? At this point we can give an answer to this question: first of all beanplots permit the handling of large datasets, without deciding the number of the data features to impose on data (in particular, bins).

At the same time beanplots can be used in two distinct ways, firstly in an exploratory way to describe the underlying data structure (Drago Scepi 2009 [236]), in particular beanplots tend to contain the information of the interval-value data, and the boxplot-value data. Secondly, beanplots can be used for the contained density trace which can be a useful tool in the analysis. It is possible in this sense to analyse and to forecast the intra-day dynamics without imposing a strong a-priori

hypothesis on the number of bins. At the same time the possibility to visualize the observations can allow the quick identification of the outliers (a mechanism that is not permitted by using different data typologies). Finally beanplots allow us to understand either the structural changes due to breaks in the original time series or the changes due to events to the series that create a change in the beanplot shape.

The kernel density estimators can be compared with other non parametric methods of density estimation (see for example Fryer 1977 [284]). Empirical results can show that, for example, the splines (see Ahlberg Nilson Walsh 1967 [8]) smooth out the original data. This implied the loss of some relevant data features. Therefore, kernel density estimators seem very useful in explorative contexts while spline smoothers retain the very relevant data features, not taking into account some irregularities which arise however, in this case of complex data such as high frequency data.

The densities present characteristics which define well the structure of the data. In particular this structure can represent well at the same time either the long run dynamics (the location of the data) or the intra-temporal variation (the size) that could be related in some specific phenomena to the risk at time t . In general these types of data represent the single observation in a symbolic data table, where the same symbolic data table can contain different types of data.

It is very important to observe the points of maxima density over time t , in fact, these points represent the "equilibrium" levels for each temporal observation at t . These equilibrium points represent valuable information that is unknown or latent, related for example to the occurrence of short term cycles and seasonalities.

6.4 Beanplot Time Series (BTS)

Definition 4. A Time Series Beanplot $\{b_{Y_t}\} t = 1...T$ is an ordered sequence of beanplots or densities over the time.

Beanplots can be viewed as time series where they are realizations of a X beanplot variable over the time t .

So at each time t we have different realizations of the interval-value data in the beanplot with the upper bound a_{U_t} and the lower bound a_{L_t} the beanline a_{C_t} and etc. So we obtain for each t the beanplot realized stylized features for the location and the size:

$$[a_{U_1}; a_{C_1}; a_{L_1}], [a_{U_2}; a_{C_2}; a_{L_2}] \dots [a_{U_t}; a_{C_t}; a_{L_t}] \quad (6.7)$$

At the same time we obtain the descriptors for the beanlines over time. The beanlines represent the location of the complex object. So we have:

$$[a_{M_1}], [a_{M_2}] \dots [a_{M_t}] \quad (6.8)$$

At the same time, similarly to an interval we can obtain the radii and the center for the beanplot. In that sense we obtain the description of the dynamic of the size over time.

$$[a_{C_1}; a_{R_1}], [a_{C_2}; a_{R_2}] \dots [a_{C_t}; a_{R_t}] \quad (6.9)$$

A very important piece of information that needs to be provided is the information on the first and the last observations in the temporal interval. When the closing value is lower than the opening value it means that the original series is falling over time. We account also for these descriptors over time.

$$[a_{OP_1}; a_{CL_1}], [a_{OP_2}; a_{CL_2}] \dots [a_{OP_t}; a_{CL_t}] \quad (6.10)$$

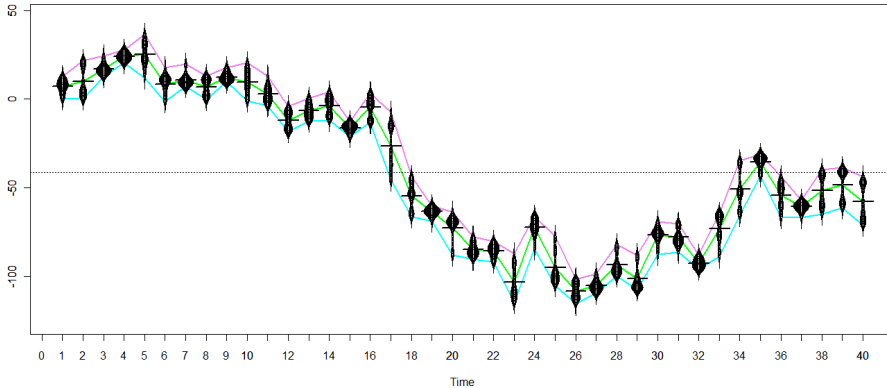


Figure 6.6: Simulated beanplot time series (BTS) and turning point identification

By observing the descriptors over time, it is possible to observe the turning points of the series over time. In particular we are interested in the local minima and the local maxima. Identification of the turning points for the three attribute time series can be very important to understand the general dynamics over the time of the beanplot time series (BTS).

It is necessary, as well, to test the attribute time series using the Turning Point test, the Difference-Sign test and the Rank test²⁷.

It is possible to define a turning point test for the attribute time series of the beanplots: given a time series of attributes $a_1 \dots a_T$ or descriptors of the beanplots $d_1 \dots d_t$ we have for the generic attribute time series a_t :

²⁷See for example Brockwell and Davis 2002 [113] and Di Fonzo Lisi 2005 [221]

$$a_{t-1} < a_t \text{ and } a_t > a_{t+1} \quad (6.11)$$

We obtain superior turning point, where the inferior one is:

$$a_{t-1} > a_t \text{ and } a_t < a_{t+1} \quad (6.12)$$

So there is a turning point if:

$$(a_t - a_{t-1})(a_{t+1} - a_t) < 0 \quad (6.13)$$

The turning point test verifies if the observed series behave as random²⁸. In that sense it is possible to show that:

$$tp_n = \frac{\hat{p}_n - 2(n-2)/3}{\sqrt{16n-29}/90} \quad (6.14)$$

Where \hat{p}_n are the observed turning points. It is possible to show that this value for $n \leq 25$ tends to distribute as a standardized normal. For the Difference Sign Test we check the different attribute time series:

Given \hat{d}_n as the number of differences $(a_t - a_{t-1})$. In the case $E(D_n)$ on the hypothesis of random converge to:

$$td_n = \frac{\hat{d}_n - (n+1)/2}{\sqrt{(n+1)/12}} \quad (6.15)$$

Another relevant test in that sense is the Rank Test²⁹.

It is possible also to consider the methods of analysis of the structural changes in the attribute time series a_t considered. It is interesting to note that structural change in the original time series corresponds to the structural change in one of these measures (or all together) as

²⁸Brockwell and Davis 2002 [113] and Di Fonso Lisi 2005 [184]

²⁹Hallin and Puri 1992 [332] for a survey of the topic

minima, maxima and centre.

Here we present the algorithm 2 for the classical analysis of the beanplot time series (BTS) and the algorithm 3 for the analysis of the structural change (see Zeileis et al. 2003 [713]).

In this sense, we can hypothesize to build different symbolic data tables using different intervals in which we compute densities over time to discover these equilibrium points over the time (that could be useful, for example in finance for trading purposes). A specific density oriented type of data is the beanplot data [416], that considers jointly some aspects of the underlying data (the value of the single data), as a specific stripchart diagram and the density in a form of density trace. A very interesting visualization is the data from the Dow Jones index computed for the period of the financial crisis, in figure 6.7 using beanplots data:

The stripchart diagram is very useful to detect the different intra period patterns in data. These different patterns can be due to seasonalities or intra-period cycles.

In this sense the choice of the length of the single temporal interval t (day, month, year) is very important and depends on the specific data features (the length of the cycles) and on the objectives the analyst wants to study (Drago and Scepi 2009).

Another important point to consider in the beanplot time series (BTS) is the possibility of taking into account the first and the last observation of the beanplots to observe the intra-temporal dynamics (increasing values over time or not).

In figure 6.7 is the beanplot time series (BTS) for the Dow Jones Market for the period 1996-2010, whilst in figure 6.8, figure 6.9 and figure 7.0 there are the enhanced beanplot time series (BTS) considering the first and the last observation for simulated and real data (Dow Jones Market and FTSEMIB Italian Market).

Data: A scalar time series Y_t and the associated beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$. Each beanplot is denoted b , the entire set of beanplots is B

Result: A beanplot time series (BTS) and a classical analysis of the attributes minima, maxima and centers

```

begin
    Choice of the interval considered  $I$ 
    Choice of the kernel  $Ke$ 
    for  $b \in B$  do
        | Computing the optimal bandwidth (Sheather-Jones
        | method) of the object  $b$ 
    end
    for  $b \in B$  do
        | Computing the  $mi$  descriptors of the object  $b$  (minima)
    end
    for  $b \in B$  do
        | Computing the  $ma$  descriptors of the object  $b$  (maxima)
    end
    for  $a \in A$  attributes do
        | Computing the trend for the  $a$  attribute of the object  $b$ 
        | Computing model selection statistics as the  $R^2$  adjusted
        | Is the trend the best approximation?
        if the trend is not the best approximation then
            | compute another order of the polynomial
        end
    end
end

```

Algorithm 2: Classical Analysis of a beanplot time series

Data: A scalar time series Y_t and the associated beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$. Each beanplot is denoted b , the entire set of beanplots is B

Result: A list of the structural change for the d beanplot time series (BTS) descriptors minima, maxima and centers

begin

Choice of the interval considered I

Choice of the kernel Ke

for $b \in B$ **do**

Computing the optimal bandwidth h (Sheather-Jones method) of the object b

end

for $b \in B$ **do**

Computing the mi descriptors of the object b (minima)

end

for $b \in B$ **do**

Computing the ma descriptors of the object b (maxima)

end

for $b \in B$ *descriptors* **do**

Compute the structural change

Test the structural change

end

end

Algorithm 3: Structural Change in beanplot time series (BTS)

6.4. Beanplot Time Series (BTS)

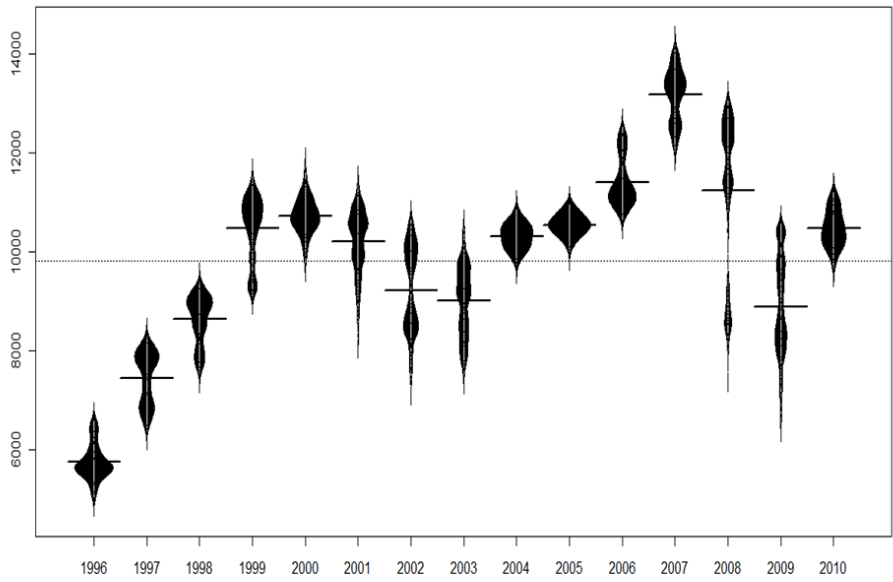
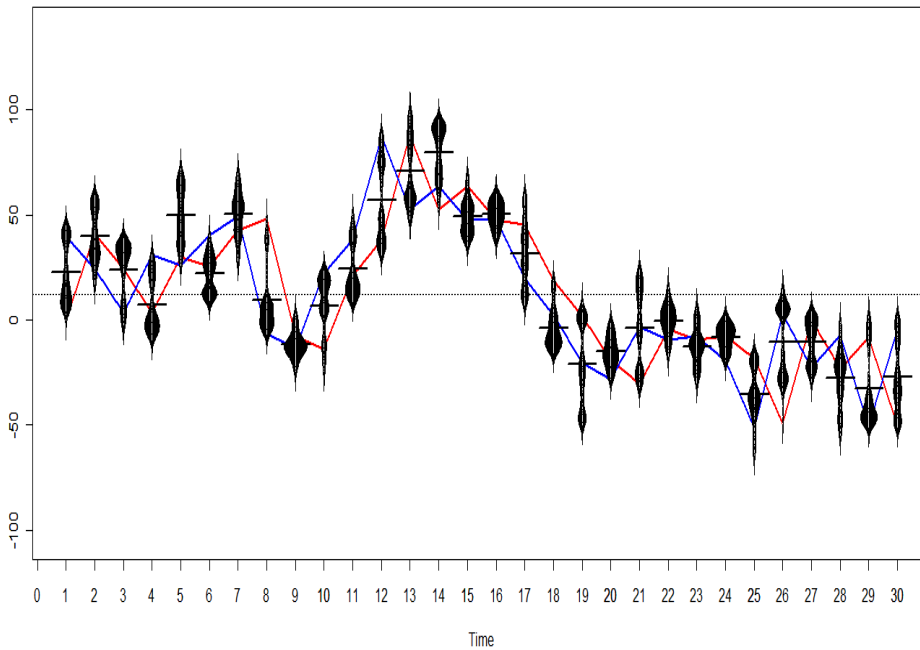


Figure 6.7: Dow Jones Index Beanplot Time Series (BTS) considered for the period 1996-2010

Figure 6.8: Enhanced density data with first and last observation (in red and blue respectively)



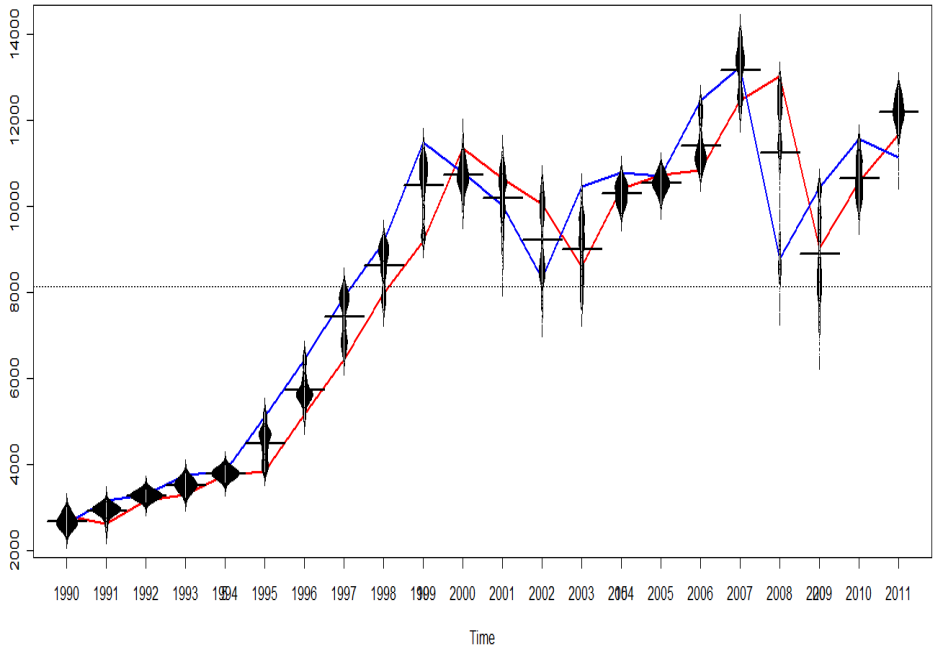
6.4.1 Beanplot Time Series (BTS): Kernel and the Bandwidth Choice

An important problem needs to be considered in the beanplot time series (BTS), those of the Kernel and the Bandwidth for each beanplot data or for the entire time series.

As we have seen the Beanplot time series (BTS) $\{b_{Y_t}\}_{t=1 \dots T}$ is an ordered sequence of beanplots or densities over time. The time

6.4. Beanplot Time Series (BTS)

Figure 6.9: Enhanced density data with first and last observation: DJI 1990-2011 (in red and blue respectively)

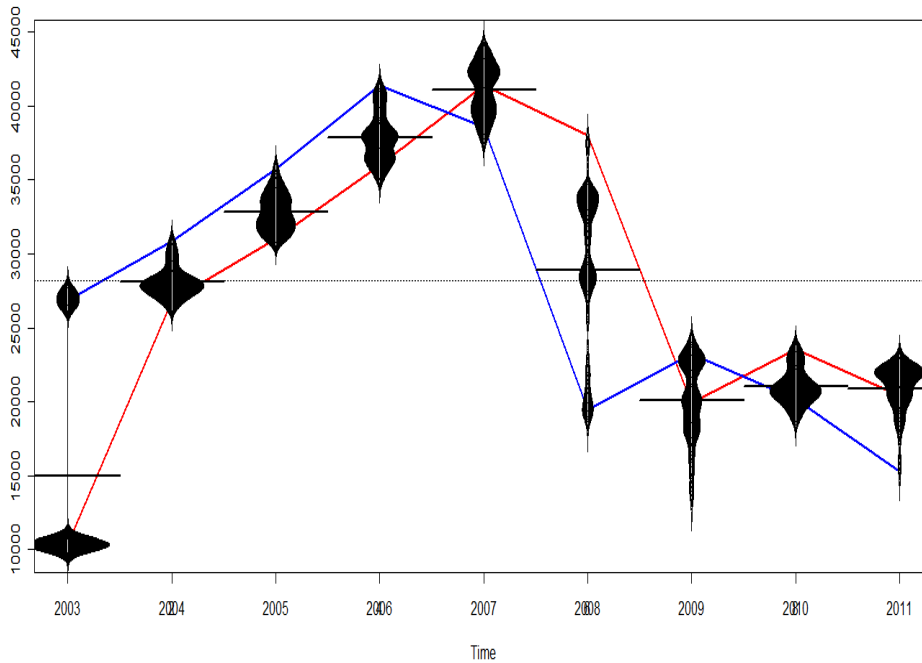


series values can be viewed as realizations of an X beanplot variable in the temporal space T , where t represents the single time interval. The choice of the length of the single time interval t (day, month, year) depends on the specific data features and objectives which the analyst wants to study.

A beanplot realization at time t is a combination between a 1-d scatterplot and a density trace.

It is possible to use various Kernel functions: uniform, triangle,

Figure 6.10: Enhanced density data with first and last observation:
FTSEMIB.MI 2003-2011 (in red and blue respectively)



epanechnikov, quartic (biweight), tricube (triweight), gaussian and cosine. The choice of the kernel in the beanplot time series (BTS) is not particularly relevant because our simulations show that the different kernels tend to fit similarly the underlying phenomena. Some differences reveal themselves in the presence of outliers. In these cases a better kernel seems to be the Gaussian kernel which is more robust. So by considering the data characteristics, we have chosen this kernel for the different applications.

The choice of the h value is much more important than the choice of K (see Silverman 1986) [633]. A large literature exists on bandwidth selection³⁰. With small values of h , the estimate looks "wiggly" and spurious features are shown. On the contrary, high values of h give a too smooth or too biased estimate and it may not reveal structural features, as for example the bimodality of the underlying density. In any case it is difficult to give a clear indication of which bandwidth selection method is the best.

With a visualization aim, we use the Sheather-Jones criteria (Sheather Jones 1991 [624]) that defines the optimal h in a data-driven approach in our application.

Here it is important to complete the part related to the beanplot time series (BTS) by focusing on the h choice of the same time series³¹. In fact it is important to note that where it is necessary to consider a beanplot time series (BTS) both for clustering or forecasting purposes it is necessary to fix one bandwidth for all beanplots (or for all the beanplot time series BTS).

For visualization and strict data exploratory purposes we can use a general bandwidth selection (the Sheather Jones method) for each beanplot data, but for other purposes it could be necessary to use another approach.

So we start from the original beanplot and we obtain the bandwidth with the usual methods. The sequence of the values of the single bandwidth represents the level of variation of the beanplot. The value that could be chosen could be rationally chosen near the median by running the algorithm 4 that visualizes the different beanplot time series

³⁰In the multivariate case the research is less developed, see for example Zhang King Hyndman 2004 [719] (see as surveys Marron 1987 [487]; Chiu 1991 [139]; Jones, Marron and Sheather 1996 [413])

³¹This aspect is clearly not existent in the literature of interval time series (ITS), where it needs to be explored more in relation to the choice of the bins in the literature of histogram time series (HTS)

(BTS) associated with the same algorithm.

In any case the bandwidth of the different beanplot visualized over time represents a useful indicator of the original data.

6.4.2 Trends, Cycles and Seasonalities

Definition 5. Beanplot Time Series (BTS) primary attributes are realizations of a single beanplot $\{b_{Y_t}\} t = 1 \dots T$ features over the time.

Primary attributes are related to the location and size of a beanplot. So we assume the $a_{U_1}, a_{C_1}, a_{L_1}$ to be composed of the components:

$$a_t = Tr_t + Cy_t + Se_t + Ac_t + U_t \quad (6.16)$$

Where Tr is a trend component, Cy is a cycle, Se is a seasonality, and Ac is an accidental component due to shocks, where U is a residual.

In this case we are assuming the mode:

$$a_t = f(t) + e_t \quad (6.17)$$

for each attribute time series. Later we will consider also for time series, beanplot descriptors d_t (or coefficients estimation).

Infact, by considering separately the $a_{C_t}, a_{L_t}, a_{U_t}$, we can compute the trend for each attribute time series, for example assuming the function $f(t)$ as a polynomial (Di Fonzo Lisi 2005):

$$f(t) = \delta_0 + \delta_1 t + \dots + \delta_q t^q \quad (6.18)$$

So we can estimate for each attribute time series $a_{U_1}, a_{C_1}, a_{L_1}$ as y_t :

$$a_t = \delta_0 + \delta_1 t + \dots + \delta_q t^q + \varepsilon_t \quad t = 1 \dots T \quad (6.19)$$

We can express the 6.19 in this form:

Data: A scalar time series Y_t and the associated beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$. Each beanplot is denoted b_t , the entire set of beanplots is B

Result: a sequence of bandwidths $h \in H$ related to the beanplot B_t over time

```

begin
    Choice of the interval considered  $I$ 
    Choice of the kernel  $K$ 
    for  $b \in B$  do
        | Computing the optimal bandwidth (Sheather-Jones
        | method) of the object  $b$ 
    end
    Is it possible to use a criteria?
    if a criteria could be identified then
        | compute the bandwidth  $h$  for the beanplot time series
        | (BTS)
    end

    Obtain a range of candidates of best bandwidth  $h$ 
    (considering eventually the median)
    for  $h \in H$  bandwidths do
        | Visualize the bandwidth for each beanplot  $b$ 
        | Visualize the sequence of bandwidths
    end
end

```

Algorithm 4: Choosing the optimal bandwidth

$$a_t = D\delta + \varepsilon \quad (6.20)$$

Where D is a matrix

With:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} 1 & t_{11} & t_{21} & \dots & t_{m1} \\ 1 & t_{12} & t_{22} & \dots & t_{m2} \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & t_{1n} & t_{2n} & \dots & t_{mn} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \vdots \\ \delta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (6.21)$$

$$y = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{bmatrix}, \alpha = \begin{bmatrix} \delta_0 \\ \delta_1 \\ \dots \\ \delta_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix} \quad (6.22)$$

So by the Ordinary Least Squares we can obtain the estimates for the parameters $\delta_0, \delta_1, \dots, \delta_q$:

$$\hat{\delta} = (D'D)^{-1}D'y \quad (6.23)$$

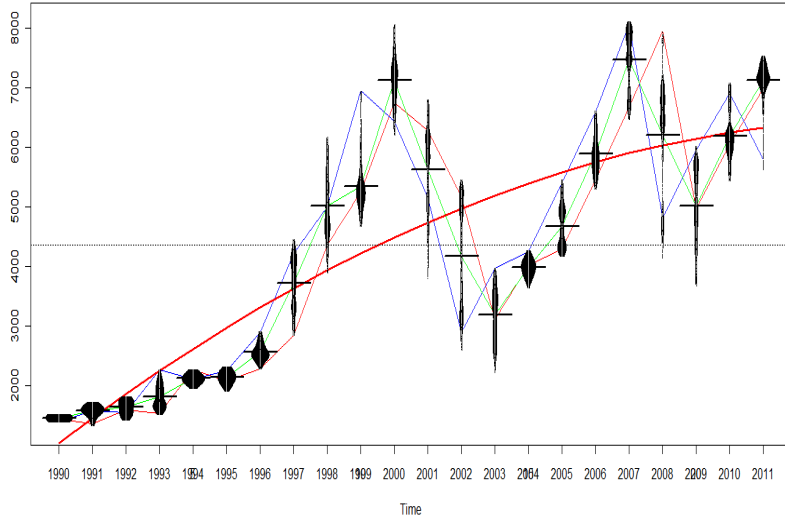
General rules of the scalar time series can be applied to select the model for the trend (in particular the maximization of the R^2) for each attribute time series. A case of trend inadequacy is represented in the figure 6.11. In the case represented in the figure it is necessary to use another strategy in the trend estimation.

Before to consider the analysis of the cycle, it is interesting to note that the beanplot time series (BTS) is a good tool for the analysis of the business cycle³², in fact they show the variability comparing the

³²In particular they seem useful to be added to the classical tools of the Business Cycle Analysis see Cipolletta (1992) [143]

6.4. Beanplot Time Series (BTS)

Figure 6.11: Enhanced Beanplot Trend for the centre: DAX 1990-2011 (in red and blue, green respectively open, close, centre)

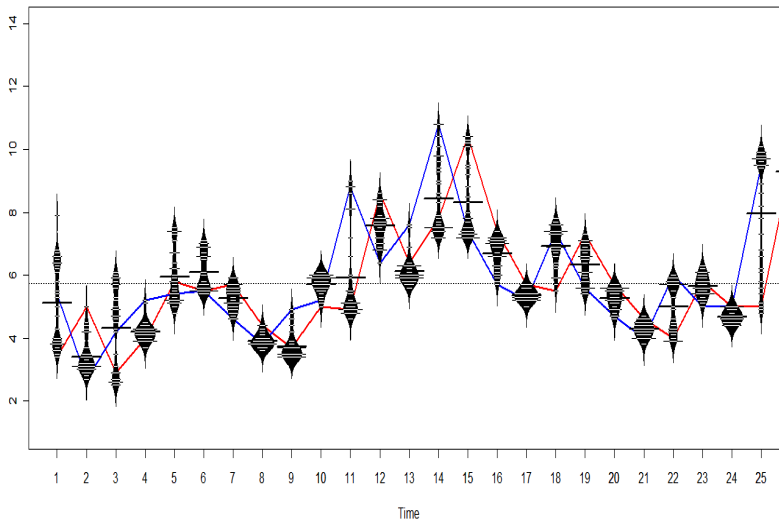


economic performances period by period. We identify in figure 6.13 the critical period in the US Total Capacity Utilization for all industry in 1975, 1983, 2009 related to the big recessions³³. At the same time in figure 6.12 we can observe the enhanced beanplot time series (BTS) and the business cycle analysis for the US unemployment rates 1948-2011. It is important to note that the visualization of the phenomena was simplified by the number of observations. In fact in using the beanplot data we are obliged to use a higher quantity of data.

Finally we can estimate the cycle of the series and also the seasonality. To estimate the seasonality for each attribute time series we can

³³Data are from Federal Reserve of Saint Louis FRED [750]

Figure 6.12: Enhanced Beanplots and Business Cycle Analysis: 3-Year US Unemployment Rates 1948-2011 (in red and blue, first and last observation)



use a set of dummy variables and estimate jointly the trend and the seasonality:

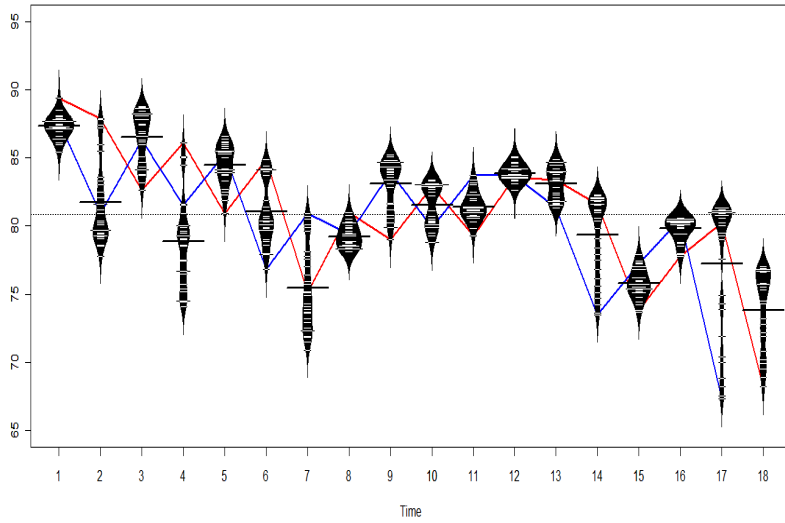
$$a_t = D\alpha + V\gamma + \varepsilon_t \quad (6.24)$$

Where V is a matrix of dummy variables related to the seasonality and the γ is a vector of coefficients that measure the impact of the seasonality. Finally O can be considered as some external shocks imposed to the time series that can be obtained by adding some specific dummy variables in the estimation. So we have:

$$a_t = D\alpha + V\gamma + O\zeta + \varepsilon_t \quad (6.25)$$

6.4. Beanplot Time Series (BTS)

Figure 6.13: Enhanced Beanplot and Business Cycle Analysis: 3-Year US Capacity Utilization: Total Industry (TCU) 1967-2011 (in red and blue, first and last observation)



The beanplot time series (BTS) shows the complex structure of the underlying phenomena by representing jointly the data location (the beanline) the size (the interval minimum and maximum) and the shape (the density trace) over the time. See figure 1, for an example of the beanplot time series (BTS). In particular the bumps represent the value of maximum density, and they can show important equilibrium values reached in a single temporal interval (and they can be used, for example, for trading purposes). Bumps can also show the intra-period patterns over the time, and in general the shape of the beanplot shows the intra-period dynamics.

The beanlines allow the computation of the trend for the Beanplot

time series (BTS). By choosing a suitable temporal interval it is possible to visualize, as well, intra period seasonality patterns. More in general, the beanplots seem to preserve the structure of the time series, but show additional relevant patterns in data, for example by showing bumps (or equilibrium levels over the time). Another important reason in using the beanplot is that these types of data can show long-run structures where they can summarize a high quantity of data over the time.

With respect to other complex objects or symbolic used in literature, beanplots data are free to show the empirical structure for each temporal interval. At the same time we obtain a smoothed visualization of the underlying phenomena. Histograms and beanplots seem complementary: where histograms can be usefully compared, beanplots tend to show the data structure, and they can show for example observation that could be considered as outliers in a time series. Box-plot can be useful to detect and to identify outliers. In applications: histograms can be useful in setting trading systems whereas beanplots seem to be very useful in risk management to analyze the occurrences of financial crashes. In each case, it is easy to provide a transformation of the beanplot into other symbolic data. For example, it is very simple to transform a beanplot time series (BTS) into an interval-valued time series (ITS).

6.5 Exploratory Data Analysis of Beanplot Time Series (BTS)

Following Shumway and Stoffer 2011 [629], by considering the beanplot extrema we are interested in exploring the data structure. We are interested, in analysing, without any particular structural imposition the dynamics over time of the beanplot structures, where they can

6.5. Exploratory Data Analysis of Beanplot Time Series (BTS)

show us the long run dynamics of both location, size and shape of the complex objects considered.

A moving average smoother m_t can be useful for understanding and for dividing the structure from the noise:

$$m_t = \sum_{j=-k}^k \alpha_j a_{t-j} \quad (6.26)$$

Where: $\alpha_j = \alpha_{-j} \geq 0$ and $\sum_{j=-k}^k \alpha_j = 1$

In general it is used:

$$m_t = \sum_{j=-k}^k \frac{a_{t-j}}{2k+1} \quad (6.27)$$

Where $w_j = \frac{1}{(2k+1)}$

In particular it is possible to average the observations using the Kernel smoothing as a moving average smoother with a weight function, or kernel. In particular we have:

$$\hat{f}_t = \sum_{i=1}^n w_i(t) a_i \quad (6.28)$$

Where it is possible to consider:

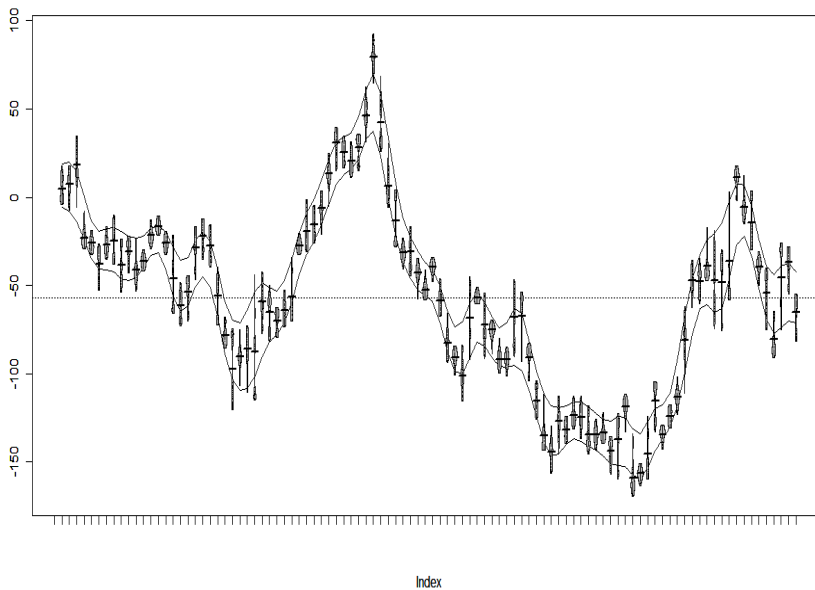
$$w_i(t) = K \frac{\left(\frac{t-i}{b}\right)}{\sum_{j=1}^n K\left(\frac{t-j}{b}\right)} \quad (6.29)$$

In which w_i are specifically the weights, and $K()$ is a kernel function. It is possible to use the normal kernel:

$$K(z) = \frac{1}{\sqrt{2\pi} \exp(-z^2/2)} \quad (6.30)$$

The Higher the bandwidth b the smoother is the result. References are Nadaraya (1964) [526] Watson (1964) [694]

Figure 6.14: Simulated Beanplot Time Series (BTS) and Kernel Smoothers



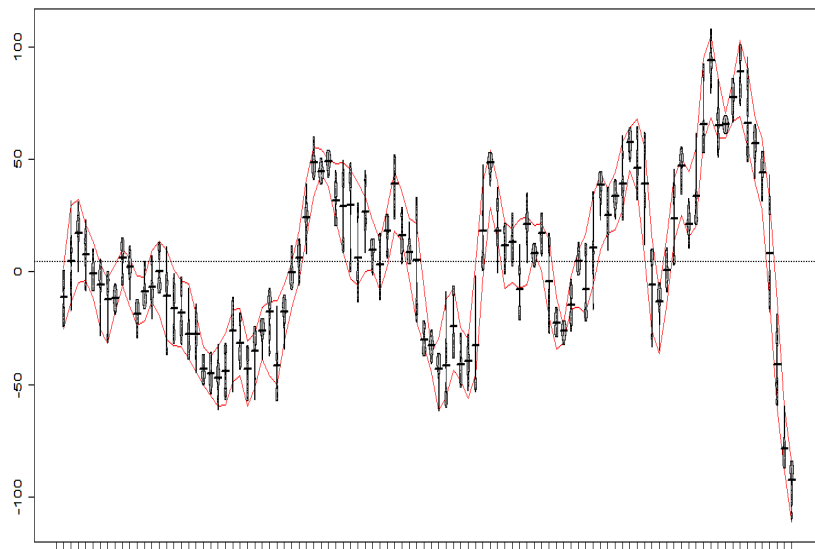
At the same time another useful approach is the use of the Smoothing splines, in which we have

$$\sum_{t=1}^n [a_t - f_t]^2 + \lambda \int (f_t'')^2 dt \quad (6.31)$$

The degree of smoothness is given by $\lambda > 0$. f_t is a cubic spline with a knot in each t (see Shumway and Stoffer 2011 [629])

See for more details Chambers and Hastie 1992 [129] Green Silverman 1994 [325] and Hastie Tibshirani 1990 [347]

Figure 6.15: Simulated Beanplot Time Series (BTS) and Smoothing Splines



These can be particularly useful in analysing financial time series to explore useful patterns (for example, for trading purposes), see in this sense Lo Mamaysky Wang 2000 [467]. Examples of the use of these techniques are in figure 6.14 (simulated beanplot time series BTS and kernel smoothers) and figure 6.15 (simulated beanplot time series BTS and smoothing splines). The complete algorithm is algorithm 5.

6.6 Rolling Beanplot Analysis

We have considered and analysed Beanplot Time Series (BTS) by starting from the original scalar time series. In all these cases the win-

now considered was specifically related to a specific temporal interval defined by the original data (as an aggregated data). In particular original data could be daily, weekly, monthly, yearly etc. The act of considering a lower frequency can be useful for many important reasons specifically linked to the analysis to be carried out. For example the risk analysis in financial data can be conducted on lower frequencies by taking into account a higher quantity of information, by considering a higher number of data available (in that case, the way to model risk is better).

Here, we consider Beanplot in a priori defined windows, where the windows change over time by adding new data, in real time³⁴. We define this type of analysis Beanplot Rolling Analysis. In this sense the window considered changes over time, where the window length is fixed. From the original Beanplot definition, we have (also Kampstra 2008 [416]):

$$\hat{f}_{h,t} = \frac{1}{nh} \sum_{i=1}^w K\left(\frac{x - x_i}{h}\right) \quad (6.32)$$

where x_i $i = 1 \dots w$ is the single observation in each t , K is a Kernel and a h is a smoothing parameter defined as a bandwidth. The window $i = 1 \dots$ varies over time by adding a new observation.

There are in this sense three approaches in the Beanplot Rolling Analysis which could be differently considered:

1. An overlapping approach (the data adjust observation by observation)
2. A non overlapping approach (the windows are completely separated)

³⁴These techniques are particularly relevant in a context where the financial data are particularly volatile and the relationships can change over time (see Pesaran Timmermann 2004 [558])

3. A partially overlapping approach (the data adjust by groups of observations)

Rolling Beanplot analysis is particularly useful in detecting the change in data structure over time. A clearly important point here is to detect a way to optimally consider the best temporal window. As will be seen later, the optimal strategy is to define the optimal window, using a number of observations of more than 30, to compute the density trace and to consider the cycle and choose the window in such a way as to not hide the cycle of the original series.

6.7 Beanplot Time Series (BTS) and Data Visualization: a Simulation Study

In order to study the performance of beanplot time series (BTS) in visualizing and exploring high frequency financial data we conduct several experiments on different models (Algorithm 6). The experiments are designed to replicate different volatility processes (with increasing complexity). In this respect we study the capability of different aggregated time series (boxplot time series BoTS and beanplot time series BTS) to capture the main features of the original data.

18 types of models (GARCH/APARCH characterized by the most simple to the most complex different volatility structures). 10 replications for each model. Each model contains characteristics of the financial time series such as volatility clusters, structural change, etc. 200,000–700,000 observations aggregated in each beanplot temporal observation. Comparing performances of different objects: Clustering (Hierarchical and PVclust) on the results, only with the descriptive aim of comparing groups of objects

The Simulation Study Design algorithm: second stage

Data: A scalar time series $\{y_t\} t = 1...T$. Each beanplot is denoted b , the entire set of beanplots is B .

Result: A list of trajectories obtained by the Kernel Ke

begin

Choice of the interval considered I

Choice of the kernel Ke

for $b \in B$ **do**

Computing the optimal bandwidth (Sheather-Jones method) of the object k

end

for $b \in B$ **do**

Computing the mi descriptors of the object k (minima)

end

for $b \in B$ **do**

Computing the ma descriptors of the object k (maxima)

end

for $d \in D$ *descriptors* **do**

Compute the kernel smoothing of the trajectories

Find the optima bandwidth of the kernel smoothing

end

end

for $d \in D$ *descriptors* **do**

Compute the smoothing splines of the trajectories

Find the optimal λ parameter

end

Algorithm 5: Exploratory data analysis and beanplot time series (BTS)

Data: A set of scalar time series $y_t\} t = 1...T$ each one representing a different model i in a set of experimented objects K . For each time series generated there is associated one specific numerical seed

Result: A set of visualization of time series of objects K , a set of clusters $p1$ and $p2$ for each time series of objects K

begin

 Choice of the interval considered I

 Choice of the number of observations n to consider

 Is it possible to compute the objects?

if *the objects cannot be computed* **then**

 | change the data structure

end

for $k \in K$ **do**

 | Computing the time series of the object k

 | Visualize the time series of the objects $k = 1...n$

end

 Is it the second stage of the experiment?

if *the second stage of the experiment* **then**

 | experiment only the best objects considered

end

for $k \in K$ **do**

 | Computing the time series of the object k

 | Visualizing the time series of the objects $k = 1...n$

 | Parameterizing the time series of the objects $k = 1...n$

 | Clustering the temporal objects using hierarchical clustering

 | Defining $p1$ clusters (1)

 | Clustering the temporal objects using PVclust

 | Defining $p2$ clusters (2)

 | Compare the clusters obtained in (1) and (2)

end

end

Choosing the two best objects in the first stage and replicating the experiments by focusing on the two best objects. Replicating the analysis on the two best objects considering the multivariate case.

A Quick Experiment: It is possible to explore the differences between the different tools using a **rolling analysis** over the time of the different tools, by considering a simulated time series.

Computational Experiments: First Stage Results

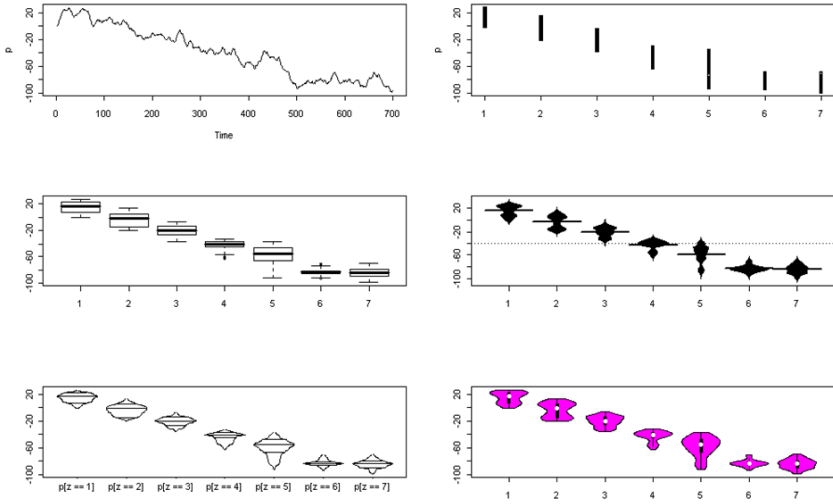
- The Scalar time series tend not to visualize **correctly** the observation in the case of high frequency data.
- Stripchart useful only for **few values** ($n \leq 30$).
- If the number of observations grows the number of the **outliers** tend to grow in the same way.
- Single observations are not visible at all in Stripchart except in the case of the outliers.
- It is very important to define the optimal **interval** in the temporal aggregation, because some features can remain **hidden**. See figure figure 6.16 and table 2

Characteristics	Stripchart	Boxplot	Beanplot	BoxPercplot	ViolinPlot
General Features	No	Yes	Yes	Yes	Yes
Trend	Yes	Yes	Yes	Yes	Yes
Cycles	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes
Structural Changes	Yes	Yes	Yes	Yes	Yes
Outliers	Yes	Yes	Yes	No	No
Intra-day Variability	No	No	Yes	Yes	Yes

Computational Experiments: Second Stage Results

6.7. Beanplot Time Series (BTS) and Data Visualization: a Simulation Study

Figure 6.16: Comparing different objects: an example on a single simulated time series: Drago and Scepi 2009

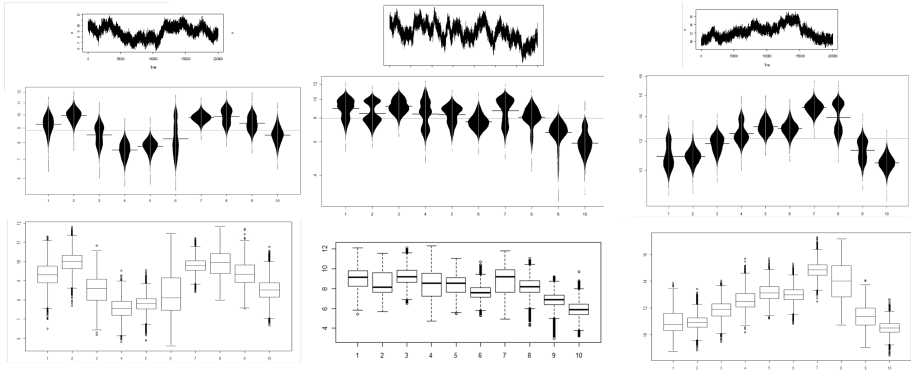


- The boxplot shows four **main features** of the temporal aggregation: **center**, **spread**, **asymmetry** and **outliers** but too many observations tend to increase the complexity of the model and increase the number of outliers (figure 6.17)
- The beanplot highlights the peaks, valleys and bumps in the distribution. Bumps are intraday price **equilibrium** levels. The number of bumps represents different market phases in the daily market structure. For asset returns, the beanplot shows volatility clusters and they can be used for the **risk** analysis (the size of beanplot actually represents a proxy of the risk).
- Beanplot becomes longer and shows price **anomalies** when there

are peculiar market behaviours (speculations).

- **Differences** between boxplots and beanplot data increase, increasing the number of observations.
- Increasing observations and increasing number of **outliers** boxplots seem to be similar. On the contrary, beanplot shows better the different intraday phases (figure 6.18)

Figure 6.17: Comparing Boxplot (BoTS) and Beanplot Time Series (BTS) (Drago Scepti 2009)[237]

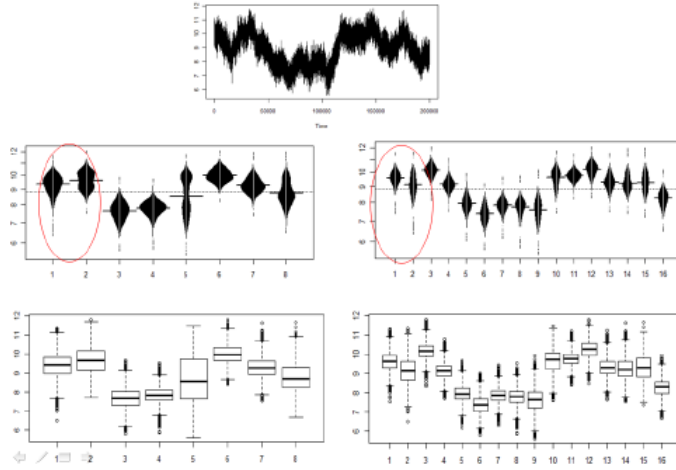


Optimal temporal window: Windows need to be directly linked to the information we are interested in, from the general trend to peculiar characteristics. Windows, in particular, need to be chosen to represent cycles exactly (if there are not other needs). Infact by choosing a higher window there could be the possibility to hide the cycle. In practice the usefulness of the beanplot is to represent complex data by removing the noise (unnecessary data characteristics) without hiding the relevant data structures, for example, the cycles. A useful type of analysis in choosing the relevant window for the beanplot data

6.7. Beanplot Time Series (BTS) and Data Visualization: a Simulation Study

could be the Spectral Analysis of a time series to define the length of the cycles (see Battaglia 2007 [66] and Hamilton 1994 [333]).

Figure 6.18: Comparing different interval temporal periods Drago Scepi 2009 [237]



For our simulations we developed several algorithms in R. We generated 18 types of models, where each model represents a different univariate GARCH/APARCH time series model

In order to analyse the effect of the different number of observations on our results, we varied, for each model, the number of observations from 200,000 to 700,000. In this way, we simulated different types of financial markets. Initially, we decided to aggregate our data in ten different groups on ten different days. Then we tested different time aggregations (by reducing or increasing the number of groups). Therefore in each day we had from 20,000 to 70,000 observations. Finally, to test our results we made 100 replications for each model.

The outcome for each computational experiment performed is the visualization of the different aggregated time series over the time. For

each experiment we registered the statistical features drawn from the beanplot time series (BTS) compared to the original scalar time series and the boxplot time series (BoTS).

The results of our simulations show in the first place that the beanplots tend to visualize a higher amount of information on the daily data, and in particular the intra-day patterns in the behaviour of the series, whereas the boxplots tend to return a smoothed view of the financial time series.

We report here, by way of example, the results with an underlying model (model 1) of the type GARCH(1,1) and those obtained with a model (model 2) of the type AR(1,5-GARCH(1,1) both with 200,000 observations.

By increasing the complexity of the time series we observe more clearly the differences between boxplot and beanplot time series (BTS). With beanplots, we are able to understand the structural changes and the different forms of the objects more clearly. When the complexity reaches a very high level there is an increase of the outliers. Boxplot time series (BoTS) seem to suffer this higher volatility of the markets. It is also interesting to note that by increasing the number of observations beanplots alone may give us clearer understanding and are therefore more useful than the boxplot time series (BoTS). In fact, the number of outliers tends to increase and the boxplots become similar to each other (in Fig. 6.17 we report an example of the model with 700,000 observations).

At the same time our simulations show that there is a specific number of observations that could be retained by choosing one interval or another. So the choice of the interval seems to be linked to the interests of the researcher. In Fig. 6.18 we show the differences between beanplot time series (BTS) with different temporal aggregations: a higher number of observations considered in the interval shows a higher number of bumps (and, of structural changes). The risk could be the loss of the information related to the cycles, where a lower number shows

the structure of the series, but it is expensive in terms of space used (and there is the risk of not visualizing patterns).

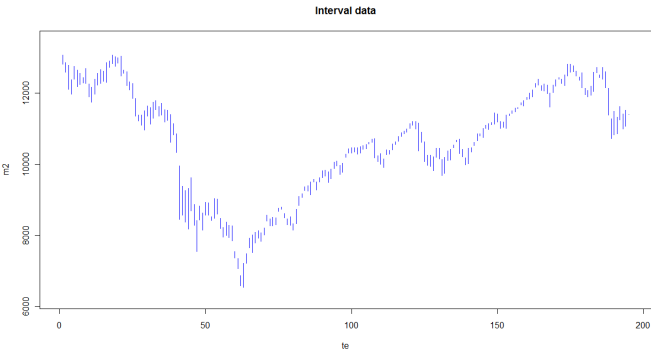
6.7.1 Some Empirical Rules of Interpretation

The aim of the analyst using the simulation is also to obtain some empirical rules of interpretation. The most important characteristic of the beanplot time series (BTS) is the capability to capture three relevant aspects in the dynamics of the complex data: the location, the size, the shape. Therefore beanplot time series (BTS) should be interpreted simultaneously considering this information.

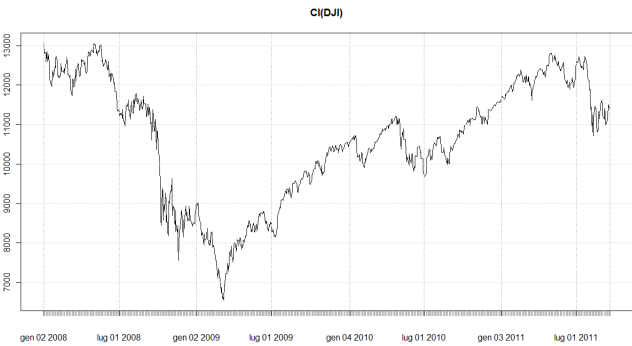
The location shows the average or median price, and thus represents a useful benchmark for comparing different units. This descriptor point gives the possibility to visualize a time series trend. This feature is not possible with other smoothers or other nonparametric techniques, while in the beanplot time series (BTS) we can explicitly consider a center for each time aggregation.

The size represents the general level of volatility, while the shape specifically represents the internal structure and the intra-day patterns. Therefore, by observing these descriptor points, we can easily identify speculative bubbles, structural changes, market crashes, etc. Furthermore, beanplot bumps can be seen as equilibrium values for the operators and they can be very important in trading strategies.

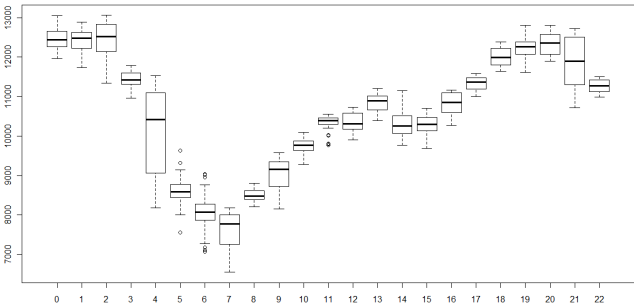
VISUALIZATION AND EXPLORATORY ANALYSIS OF BEANPLOT DATA



(a) Interval Time Series - ITS (Week)



(b) Scalar Time Series (STS)



(c) Boxplot Time Series (BoTS)

6.8 Visualization: comparing the Beanplot time series (BTS) to other approaches

An interesting experiment is now related to the use of real data (and the associated problems of the choice of the interval temporal). In particular we look at the Dow Jones scalar time series in which we consider the period related to the financial crisis 2008-2011. We consider comparatively the scalar time series (STS), the interval time series (ITS), the boxplot time series (BoTS), the histogram (HTS) one and the beanplot time series (BTS) (figure 6.18.2 and figure 6.18.3).

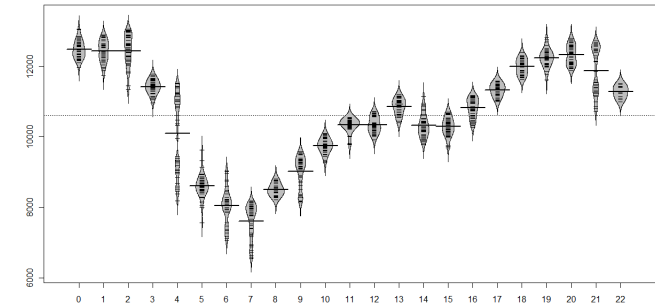
A first observation can be done in the choice of the week as a natural temporal interval for the intervals. The first impact on the visualization shows that the different objects tend to visualize correctly the most important features such as the cycles (the long run cycles in particular) and the trend. It is chosen by considering the length of the cycle (so as to not eliminate its structure) by considering some information a priori. The interval time series (ITS) show the structure of the data, but in this case the problem of the number of the data can persist. Choosing a lower temporal interval means we cannot observe the intra-period variation.

The boxplot can consider the interval (weekly), a higher temporal interval, but the problem is that it tends to smooth the original data and so not show the intra-period structural changes. At the same time the relevant information is preserved.

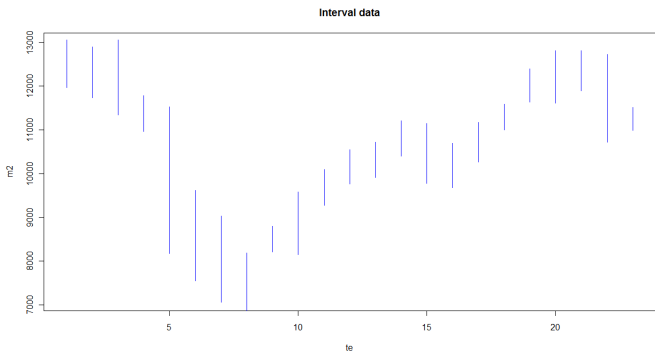
The original scalar data normally show the structure of the data, but we are unable to observe all the observations in a good way, due to the high number of data.

Using beanplot data we are able to observe the data, and in particular we are able to observe the intra-period structural changes (the bumps).

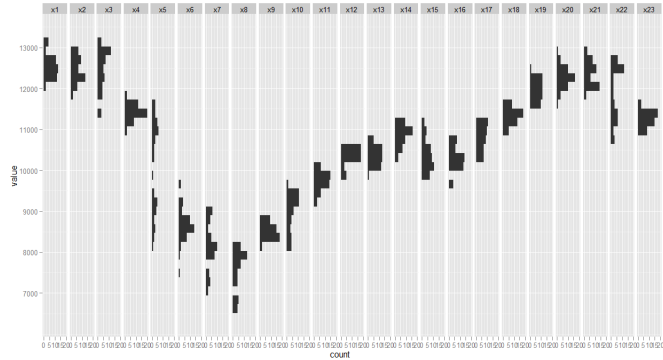
VISUALIZATION AND EXPLORATORY ANALYSIS OF BEANPLOT DATA



(d) Beanplot Time Series (BTS)



(e) Interval Time Series - ITS (2 months)



(f) Histogram Time Series (HTS)

At the same time, by considering the intervals, a lower interval can be important because in that case we are able to understand the intra-temporal structural change.

6.9 Applications on Real Data

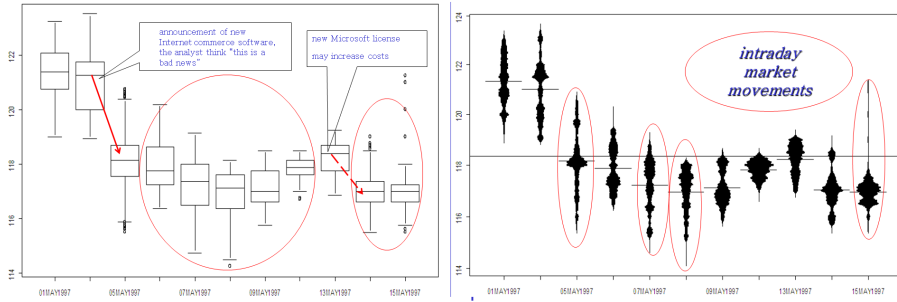
6.9.1 Analysing High Frequency Data: the Zivot dataset

The data used in this application are contained in the Zivot dataset (see Yan Zivot 2003) [710]. These data are specifically related to the official TAQ (Trades and Quotes) database containing "tick by tick" data for all the stocks on NYSE from 1993. The Zivot dataset refers to 1997 and contains quotes and the trades for Microsoft (figure 6.19). Here we consider the transaction prices for the period 1 May-15 May for a total of 11 days (except periods where the market is closed). Finally, we take into account 98,705 observations (instead of 98,724). In this case we do not consider the prices > 150 , which allows us to avoid the data visualization. This exclusion does not modify the data structure.

The conclusions for the analysis of the real time series of high frequency data are similar to the simulated one. Each beanplot represents, as in the simulated data, a day of market transactions.

Each beanplot can be seen to be the ideal "image" of the market at a specific time. In particular we can observe that the objects seem to be characterized by a response to the shocks, as the level (or the average) of the boxplots and the beanplots tends to change day by day. This phenomenon is due to the response of the time series to news that impose a different size, shape and location conditionally to the relevance of the shock. Changes in boxplot and beanplot levels seem to be directly influenced by daily news, whereas the number of

Figure 6.19: High Frequency Microsoft Data 1-15 May 2001 (see Drago and Scepi 2009)



bumps in the beanplot time series (BTS) is directly linked to intra-day news. At the same time it is interesting also to note that volatility levels seem to be higher after a single shock and tend to decrease over the time, and disappear after a few days. Finally it is important to note that the structure of the time series appears highly irregular in the beanplot case. At the same time the boxplots tend to smooth the information contained in data, whereas the beanplots tend to reflect the complex behavior of the markets and the intra-daily patterns.

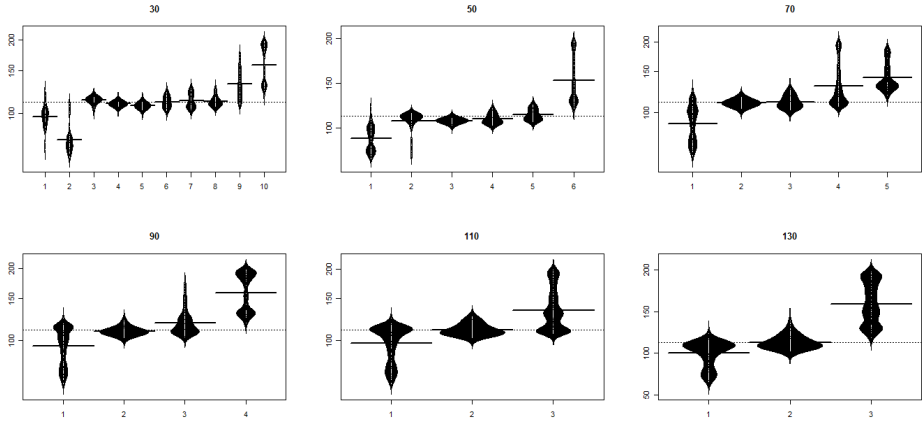
6.9.2 Application on the US Real Estate Market in 1890-2010

From the Shiller data sets (see Shiller 2005 [626]) (figure 6.20 and figure 6.21) we consider the Real Home Price Index for a long run period 1890-2010, by using the Rolling Beanplot we are able to visualize if this type of data allows the observation of the growth of the speculative bubble. So we consider various different windows (useful to compute the kernel density estimations for various subperiods) with the specific aim of observing the anomalies on prices over time. We do not find any anomaly until 2008 in which we can observe that the beanplot

6.9. Applications on Real Data

tends to show a strong change from the original series. Some values can be considered outliers, which mean that there was the growth of a speculative bubble in the period (2007-2008). The result is confirmed by the fact that applying the rolling scalar time series, the Dickey Fuller Test, we found that the series start to be non stationary in the subperiod 210-240. At the same time an analysis on the structural breaks of the period show that there is a significant structural break in the the second quarter of the year 2000.

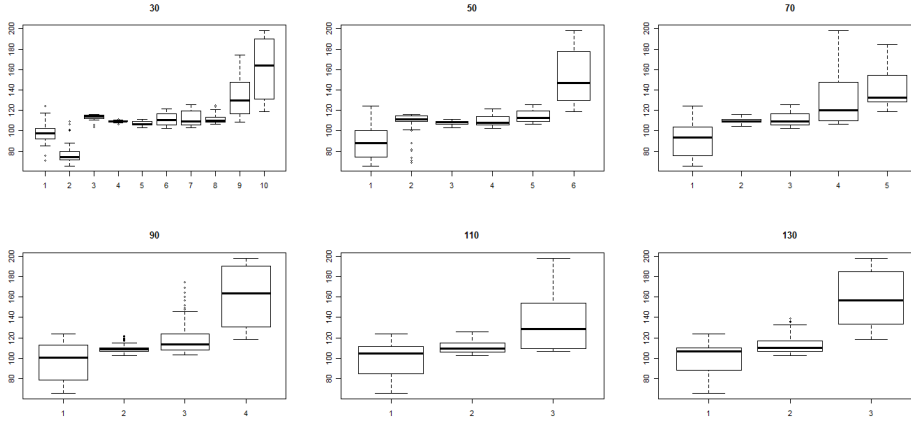
Figure 6.20: Rolling Beanplots Real Home Price Index 1890-2011 using different windows



6.9.3 Comparing Instability and Long Run Dynamics of the Financial Markets

We performed an analysis on some of the most relevant markets around the world. The period considered are the years 1990-2011, where 2011 is related to the period 1 January- 14 August. The missing data are not considered. The analysis is divided into two phases:

Figure 6.21: Rolling Boxplots Real Home Price Index 1890-2011 using different windows



1. Extracting the bandwidth for each beanplot time series (BTS) to measure the instability (table 6.1 and table 6.2)
2. Computing the trend to measure the long run growth of the beanplot time series (BTS) (table 6.3–6.6)

6.10 Visualizing Beanplot Time Series (BTS): Usefulness in Financial Applications

Various financial operations can be improved through considering the visualization of the beanplots. First of all, the beanplot is a useful tool for the monitoring of the market and in discovering speculative bubbles. In particular, it helps to monitor both short and long run dynamics. At the same time, the beanplot leads to the consideration of deviations from long run equilibrium values and quickly detect any

6.10. Visualizing Beanplot Time Series (BTS): Usefulness in Financial Applications

changes. These changes can be caused by structural factors (for example policies) which could change the beanplot structure. So, the beanplot can be useful to monitor more than one stock at a time. Thus, it is useful in a asset allocation context, where it is necessary to consider the performances of different stocks. In this sense, it is possible to mix elements of the fundamental analysis and other financial techniques to relate the outcomes from these techniques to the location, the size, and the shape of the beanplots. In this way it is possible to anticipate the impacts of financial events on the beanplots. The capability of the tool to summarize a large quantity of information could be very useful in the monitoring of a large number of stocks. Visualization is useful also in risk management techniques. In fact, it is possible with the beanplot data to monitor the evolution of the identified risks (also through using other techniques such as control charts). Here, the beanplot is useful also in the phase of scenario analysis as it is possible to consider and to use simulation methods to predict the impact over time of different decisions and their outcomes. Beanplots allow the visualization of the outcomes of many different economic policies in comparative scenarios.

Possible applications: Market Monitoring and discovering speculative bubbles, discovering financial market patterns (Statistical Arbitrage), Asset Allocation, Risk Management, Scenario Analysis.

Summary Results: Visualization
Beanplots represent a useful Internal Representation which uses Kernel Density Estimation.
Beanplot Time Series (BTS) allow the representation of High Frequency Data.
Beanplot Time Series (BTS) retain the information of the very long underlying time series.
In the Visualization process, an optimal bandwidth can be obtained by the Sheather Jones Method.
A simulation study allows the observation of the informative content of the Beanplot Time Series (BTS), with respect to other types of Internal Representations.
Real data allows the best interpretation of Beanplot Time Series (BTS) in real contexts. In particular, we can observe the volatility levels by each day, the equilibrium levels (useful in structural changes), the intra-period seasonalities, etc.

Table 6.1: Bandwidth for various beanplot time series (BTS) 2005-2011

	X000001.SS	N225	IETP	IBEX	FTSEMIB.MI
2005	22.81	196.97	11.07	113.39	339.04
2006	46.19	217.08	19.49	185.15	335.54
2007	175.91	219.05	23.63	145.35	419.83
2008	196.83	306.78	34.38	321.28	867.18
2009	113.43	216.13	13.87	240.02	557.85
2010	48.00	161.00	7.65	179.00	262.95
2011	42.50	104.35	2.38	147.93	319.63

6.10. Visualizing Beanplot Time Series (BTS): Usefulness in Financial Applications

Table 6.2: Bandwidth for various beanplot time series (BTS) 2005-2011

	GDAXI	FCHI	DJI	BVSP
2005	59.20	51.91	55.43	564.32
2006	100.72	77.15	102.33	627.24
2007	86.23	50.08	115.79	1346.42
2008	141.92	119.48	239.58	2166.38
2009	132.05	73.23	223.78	1888.33
2010	60.24	64.26	148.34	931.94
2011	51.52	37.87	132.86	1279.58

Table 6.3: BVSP

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	32230.1869	3275.0033	9.84	0.0000
poly(tt2, 3)1	52414.0124	14275.4086	3.67	0.0023
poly(tt2, 3)2	43927.9664	14275.4086	3.08	0.0077
poly(tt2, 3)3	12556.3558	14275.4086	0.88	0.3930

Table 6.4: DJI

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8209.2913	262.4511	31.28	0.0000
poly(tt2, 3)1	13723.9417	1231.0050	11.15	0.0000
poly(tt2, 3)2	-4676.9175	1231.0050	-3.80	0.0013
poly(tt2, 3)3	-938.0513	1231.0050	-0.76	0.4559

Table 6.5: GDAXI

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4275.2110	244.3792	17.49	0.0000
poly(tt2, 3)1	7518.4447	1146.2402	6.56	0.0000
poly(tt2, 3)2	-1434.3096	1146.2402	-1.25	0.2268
poly(tt2, 3)3	584.6972	1146.2402	0.51	0.6162

Table 6.6: FTSEMIB.MI

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	27303.5908	1279.8885	21.33	0.0000
poly(tt2, 3)1	-4151.3575	3839.6654	-1.08	0.3290
poly(tt2, 3)2	-19541.1928	3839.6654	-5.09	0.0038
poly(tt2, 3)3	12011.9352	3839.6654	3.13	0.0260

Chapter 7

Beanplots Modelling

In this chapter we propose a new approach for modelling time series as complex data¹, in particular financial data. For the new approach, in the same way as the works proposed by Arroyo 2009 [32] and Maté (2009) [491], the data is not scalar but is a representation of the intra-period variation as an interval, a histogram: in this case it is a density that presents interesting properties (Chapter 5). This type of data, can be clustered (Chapter 9) and forecasted (Chapter 8). The aim, in the present chapter, is that of modelling the variability of intra-period, that which relates to the temporal intervals. We define these types of models as "internal models" (see also Signoriello 2009 [630]) and "external models", which are related to temporal dynamics.

In practice the idea is to reduce the errors related to the measurement error (see Rabinovich 1995 [568]), by considering a mathematical model of the aggregated data². Original data are infact characterized by two distinct parts: the first, is structural and the second, is the

¹Here we use the Diday definition of time series as complex data: see Diday 2006 [208]

²In particular we model the original data as a density. See for mathematical modelling in this context Gershenfeld 1999 [295]

noise, which is influenced by many factors such as data incompleteness, errors, measurement errors, etc.³. Signoriello (2009) [630] also proposes this approach for histogram data⁴. In particular, the modelling approach (the internal model) is based, here, on the beanplot data (Kampstra (2008) [416]) as already presented in Chapters 5 and 6. These types of new aggregated time series can be fruitfully used when there is an overwhelming number of observations, for example, in High Frequency financial data (this is different from other types of aggregated data that do not faithfully represent the data structure).

The initial beanplot is explicitly modelled to permit the extraction of the real data structure. At the same time we have seen in Chapters 1 and 2 that there are cases in which data are overwhelming and that using some aggregation can lead to a direct loss of information, for example in financial time series. These cases also determine measurement error, where original scalar data present errors (see how high frequency financial data contains errors in Dacorogna et al. 2001 [163] and also Brownlees and Gallo 2006 [115]). In particular, high frequency financial data shows some relevant characteristics (they are inequally spaced and contain errors), which suggest the use of some alternative methods like internal representations⁵ and the modelling presented here. In this sense we have considered the use of the beanplot, or density data (Kampstra 2008 [416]) in Drago Scepi 2009 [236] that summarizes the initial data through returning the relevant data

³See in this sense the introduction to the workshop Knemo in 2006 in Naples that shows the idea of this chapter very well [1]

⁴"According to the classical theory of measure, the data generated by the "correct model" are more "real" than the empirical one, because they are purified from error sampling and from error of measurement. We should never forget that there are no "real" models, but rather models that approximate the reality in a more or less accurate manner.." Signoriello 2009 [630]

⁵Here we use the same term used by Lin Keogh Li and Lonardi 2007 [422] where the methods used are different because we represent the initial data as a density (whereas in the literature intervals, histograms etc. are used)

features. It is important to note that the beanplot time series (BTS) presented in Chapter 6, considers not only the extreme values of the interval period but examines above all the intra-day dynamics⁶. In this case it could be relevant to take into account the structural part of the model and the noise related to the different density traces (as seen in Chapter 6).

In practice we propose a transformation of the original time series into a density time series in order to analyse the variability intra-day over time where it is necessary to extract the structural part from the noise. The advantage of the approach can be seen by the fact that it retains all the relevant information of the initial data in the density plots and yet avoids the errors contained in the original data⁷. The coefficients estimation and the descriptor points will substitute the original data, and the original beanplots as well. This approach allows us to take into account not only the aggregated values of the data, but also⁸, the entire intra-day variation. The visualization of the beanplot time series (BTS) gives us the opportunity to retain all the relevant information⁹.

⁶See for different methodological approaches and examples Maté 2009 [491] and Arroyo, Espínola, Maté 2011 [36] Maia, De Carvalho, Ludermit 2008 [476] Arroyo et al. 2010 [39]

⁷It is important to stress that where the financial data are an excellent example of data that present a noise (see in that sense Sewell 2008 [619]), at the same time other data types contain errors

⁸In high frequency data the last observation could also be considered: see Dacorogna et al. 2001 [163]

⁹The information can be related to the location, size and shape of the aggregated data (see Drago Scepi 2010 [236] and Drago Scepi 2010 [237])

7.1 Beanplot Coefficients Estimation

We have seen in Chapter 6 that a beanplot time series is an ordered sequence of beanplots over the time. Each temporal interval can be considered as a domain of values that is related to the chosen temporal interval (daily, weekly and monthly). The problem introduced at this point of the thesis is this:

$$Data = Model + Noise \quad (7.1)$$

In this sense modelling is a relevant operation in the knowledge extraction process from models or by modelling¹⁰. At the same time modelling can be considered to be either based on structural distributional hypothesis or not (soft modelling)¹¹. Here, the term "Model" is equivalent to "Knowledge".

The noise is different from error in Statistics. An exploratory phase and careful pre-processing is necessary so that data can be characterized, as in the case of the high frequency data and complex data in general (figure 7.1), by structural errors, measurement errors, missing values, outliers, or in general, inconsistent values. However data are not correctly aggregated so they show some problems of aggregation loss. Last but not least, they can be structurally incomplete and they need to be integrated from a different source.

The choice of the temporal interval, also at the modelling phase, is an a priori choice and depends on the specific data features the analyst wants to study, but can also hide some important information of the data (Drago and Scepi 2009 [237]). For example, in financial analysis on the risks associated to portfolios it is usual to consider

¹⁰In the same spirit as the workshop Knemo 2006 on the Knowledge Extraction and Modeling [1] organised in September, 4th-6th 2006 at Villa Orlandi Island of Capri, Italy

¹¹More in general it could be intended as a model learned from data in the sense of Friedman Hastie Tibshirani 2009 [282]

7.1. Beanplot Coefficients Estimation

higher rather than lower temporal horizons. So, it could be considered usual to take into account higher temporal horizons (say, yearly) when looking at bad situations or crises¹². At the same time, choosing an interval or alternative temporal can change the allocation between structural part and noise and so can hide some relevant information of the models.

As we know, the beanplot can be considered a particular case of an interval-valued modal variable, like boxplots and histograms (see Arroyo and Maté (2006)). In a beanplot variable we take into consideration the intervals of minimum and maximum and the density in the form of a kernel nonparametric estimator (the density trace: see Kampstra (2008)).

The density trace is combined with a 1-d scatterplot where every single dot can be represented for each observation. The beanline can be considered to be a measure of the centre and could be represented by the mean, or the median. So the same beanplot can be considered a density trace with an interval composed of the two consecutive sub-intervals through the beanline (the radii of the beanplot).

The density trace in particular characterizes the beanplot (or the density data) and could be decomposed by the structural part and the noise.

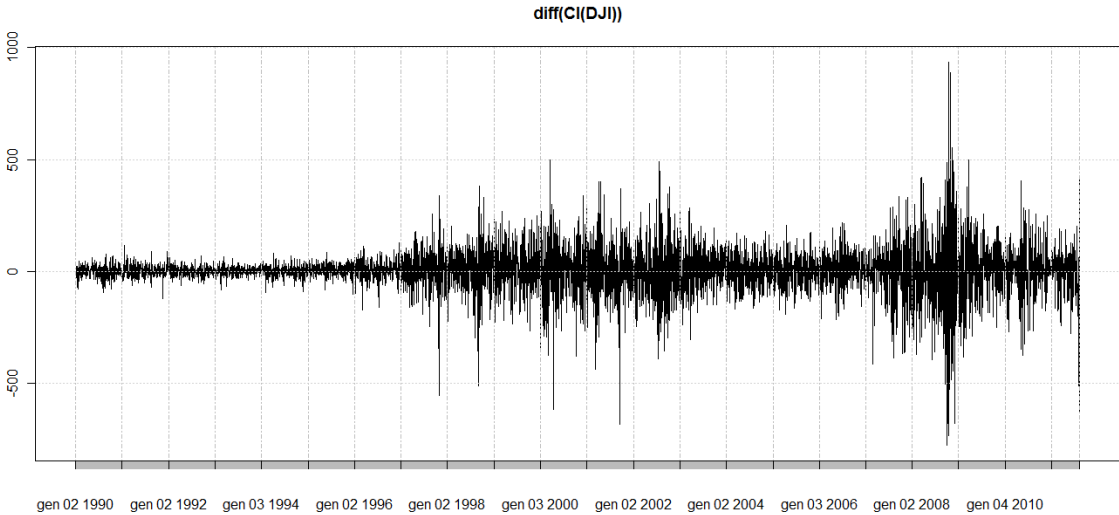
In the modelling process (Algorithm 7), more than in the simple data analysis, the choice of the h parameter is fundamental.

In fact the error term N can create irregularities that could be eliminated by a lower h . Also a lower h can hide some relevant features of our data.

In fact the higher the h parameter (the bandwidth of the density) the more irregular the curve. Therefore we need to choose carefully the parameter for the bandwidth.

¹²For the problems in risk management, real world examples and analyses using scalar data see Jorion 2006 [414] Resti Sironi 2007 [580] and Saita 2007 [601]

Figure 7.1: US Dow Jones differenced time series 1990-2011



This parameter is obtained by the Sheather-Jones method (see Kampstra (2008) [416]).

7.1.1 Beanplots Model Data: the modelling process

We start from a time series $\{y_t\}$ with $t = 1 \dots T$ an overwhelming number of observations. Our aim is to summarize the initial data by retaining the main characteristics of the original time series. A first exploratory data analysis is necessary to detect the noise in the data in accordance with the subsequent definition we have given.

7.1. Beanplot Coefficients Estimation

Data: A scalar time series y_t

Result: A coefficient estimation for each beanplot B_t in the beanplot time series (BTS)

```
begin
  Preprocessing of the original time series  $y_t$ 
  Visualization of the beanplot time series (BTS)
  for  $B \in T$  do
    Computing the  $h$  bandwidth using the Sheather Jones
    criteria
  end
  Choice of the  $I$  interval temporal unique  $t = 1 \dots T$ 
  Choice of the  $Ke$  kernel
  Choice of the  $h$  bandwidth to use
  Transforming  $y_t$  in a beanplot time series (BTS)
   $\{B_{Y_t}\} t = 1 \dots T$ 
  Is the variability represented?
  if the variability is not adequately represented then
    change the interval temporal  $I$ , the number of
    coefficients  $n$  or the bandwidth  $h$ 
  end

  for  $t \in T$  do
    Coefficient estimation of the beanplot  $B_t$ 
  end
  Is the internal model not fitting data adequately?
  if the internal model is not adequately fitted then
    change the interval temporal  $I$  number of coefficients  $n$ 
    or the bandwidth  $h$ 
  end

  The model coefficients substitute the beanplots
end
```

Algorithm 7: The internal modelling process

We identify the **outliers**¹³ in the series $\{y_t\}$, by using statistical procedures and adequate tests like the **Dixon test** for example.

In particular in high frequency time series the cleaning process of the time series is not needed. We perform, as well, the eventual transformations of the time series.

We substitute the initial data with the different subperiods represented as Beanplots Y_{b_t} with $t = 1...T$. Subsequently we visualize these subperiods of the series by means of a **beanplot time series (BTS)** considering a unique bandwidth for the B_{Y_t} (Drago and Scepi 2008).

By choosing a different interval it is possible to obtain a different beanplot structure, but the **general features of the initial time series** (trends, cycles, structural changes etc.) tend to be preserved. In general it is important to visualize adequately the original data in the form of beanplot time series (BTS) and compute the bandwidth by means of the Sheather Jones method¹⁴. Secondly we can use this information to compute a unique bandwidth for the entire series, (using the median of the bandwidths). Therefore we decide to define a proper t temporal interval (say, daily, monthly etc.) so as to aggregate our observations. We substitute the initial data with the aggregated time series s_t with $t = 1...T$. Subsequently we visualize this aggregated time series by means of a beanplot time series (BTS) B_t (Drago Scepi 2010 [237]). This allows us to represent the location (or the centre), the size and in particular the intra-period variability, which is not so manifest in the original time series.

By choosing a suitable temporal interval it is possible to visualize, as well, intra period seasonality patterns. In general, the beanplots

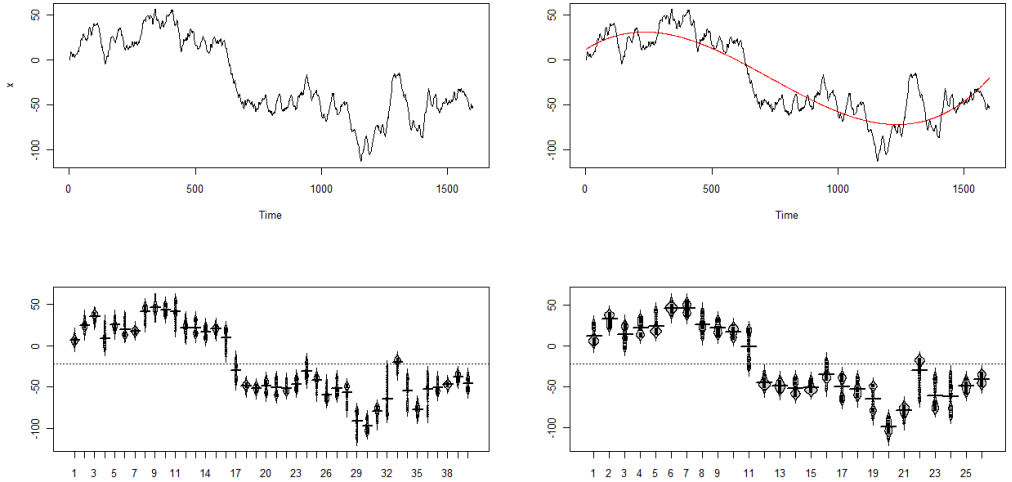
¹³ The identification of the outliers is relevant because they can have an impact on the results of the analysis based on scalar data: see Chalabi and Würtz (2009) [128]

¹⁴ That represent a curve rougher (an h higher) than another smoother one (with a lower h). See in this sense Wand and Jones 1995 [687]

7.1. Beanplot Coefficients Estimation

seem to preserve the structure of the time series, but show additional relevant patterns in data, for example by showing bumps (or equilibrium levels over the time). Beanplots using long time series can show long-run structures as they can summarize a high quantity of data over the time (and associated structural changes). Therefore we decide to define a proper t **temporal interval** (daily, monthly etc.) when considering our observations. The temporal interval must be coherent with the problem to be solved figure 7.2.

Figure 7.2: Beanplot Time Series (BTS) using different Temporal Intervals on an ARIMA(1,1,0) with a structural change



The second step is the choice of the adequate kernel and the bandwidth. It is important to note that the choice in the first part of the thesis was needed for each beanplot data to original observations, instead here we are choosing a specific kernel and a specific bandwidth

for all the observations.

For the choice of the unique bandwidth for the entire series, various methods can be used in order to define the best one in advance: Jones, Marron and Sheather (1996) [413] or also the Sheather-Jones criteria that defines the optimal h in a data-driven choice (Kampstra 2008 [416]). The mean, or the median from the different bandwidth, could be considered the bandwidth for the entire series.

By using the adequate bandwidth for the entire series we can discover the "real" data structure. Contrary to other complex objects used in literature, beanplots data leaves data free to show the empirical structure for each temporal interval, and we obtain a smooth visualization of the underlying phenomena.

Histograms and beanplots seem complementary: whereas histograms can be usefully compared, beanplots tend to show the data structure, and they can show observations that could be considered as outliers in a time series. Boxplots can be useful in detecting and identifying outliers. As well, in beanplot it is possible to detect outliers, in fact, every single observation is represented. This feature is useful to detect visually observations which are distant from the others. The beanline at time t is a location measure of the beanplot.

In applications: histograms can be useful in setting trading systems, whilst beanplots seem to be very useful in risk management for the analysis of the occurrences of financial crashes. In each case it is simple to provide a transformation from a data like the beanplot to other symbolic data. For example, it is easy to transform beanplots into an interval-valued time series (see in this sense the Chapter 5).

The next step is the coefficients estimation: here two different strategies are used to identify if the data calls for a different approach, where common to both is the diagnostics, then eventually a respecification of the models if required.

The beanplot time series (BTS), free from the Noise part, shows the complex structure of the underlying phenomena by representing

jointly the data location (the beanline) the size (the interval minimum and maximum) and the shape (the density trace) over the time. See figure 1, for a beanplot time series (BTS). In particular the bumps represent, in this case, the value of maximum density, and they can show important equilibrium values reached in a single temporal interval (for example, trading purposes). Bumps can also show the intra-period patterns over the time and more in general the beanplot shape shows the intra-period dynamics. When the beanplot increases, (hence, an increase in the difference between minimum and maximum) this can be interpreted through the presence of a structural change on the underlying time series (figure 7.1). The beanlines allow us, as we have seen, to compute the trend for the mean Beanplot time series (BTS). This is an important aspect because we can detect a general growth of the original time series. Also, the growth of the trends linked to minima and maxima shows an increase in volatility and uncertainty over time.

The chosen temporal interval allows us to represent the **location** (or the centre), the **size** and in particular the **intra-period variability** over time t , not so manifest in the original time series, so the growth over time of the beanplot models seems to indicate an increase of the uncertainty over time (Table 7.1).

7.2 Coefficients Estimation: The Mixture Models Approach

As defined in equation 1, we can see that each beanplot is related to a fixed kernel and a chosen bandwidth. The coefficient estimation is necessary for the aim of forecasting the beanplot time series (BTS). So we fix both K and h and we derive the density model time series as a mixture of different distributions. We retain the coefficients p_j

Table 7.1: Internal Representations and Descriptor Points

Data	Descriptor Points
Interval	Upper/lower bound, center/radius
Boxplot	quantiles
Quantile	quantiles
Histogram	midpoints (upper/lower bounds), counts
and densities	
Histogram: structural part	coefficients obtained by the B-splines
Beanplot	standard coordinates
Beanplot: structural part	coefficients obtained by the Mixtures

Figure 7.3: The Data Analysis Cycle

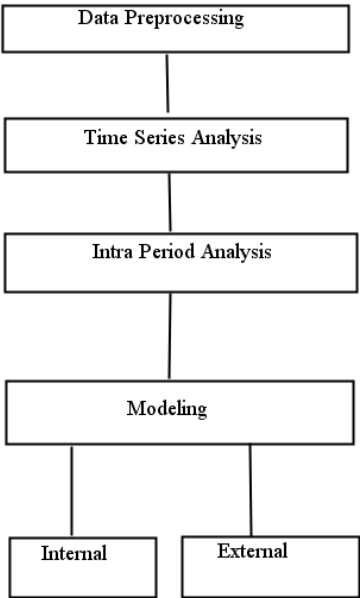
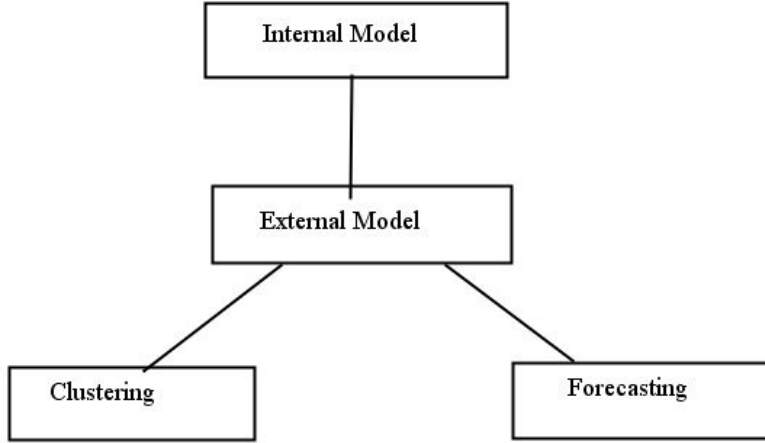


Figure 7.4: Internal and external modelling



representing the different components of the distributions occurring in the mixture. Then we decide to build t vectors of k coefficients, A_t , and substitute the beanplot time series (BTS) with it. Therefore A_t is defined as:

$$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]' \quad (7.2)$$

We also consider a measure of the goodness of fit I_t (i.e. the Bayesian Information Criteria) representing the quality of the representation. We take into account the coefficients sequences that could be statistically tested to check the structural changes over time t .

The modelling approach (figure 7.3) is necessary to reduce the measurement error from the data. The idea is to **transform** the beanplot data into **models** to control the error deriving from empirical data (Signoriello 2009, Drago Lauro and Scepi 2009).

In particular we assume that the Beanplot represents the **intra-day variability** plus a **measurement error**.

$$\text{Beanplot Data} = \text{Model} + \text{Error} \quad (7.3)$$

By considering the visualization of the beanplot time series (BTS) we choose a kernel and a bandwidth. With the aim of modelling and forecasting beanplot time series (BTS), a **coefficients estimation** becomes necessary.

We can assume each initial intra-day observation characterized in this sense:

$$g(x) = p_1 f_1(x) + \cdots + p_k f_k(x) \quad (7.4)$$

With $0 \leq p_i \leq 1$ and $i = 1 \dots k$ and also $p_1 + p_2 + \cdots + p_k = 1$. So we have $g()$ as a finite mixture density function. The coefficients $p_1 \dots p_k$ will be the model coefficients where the $f_1(.) \dots f_k(.)$ are the component densities of the mixture. So we have: $f_j(x) = f(x|\theta_j)$ where θ_j defines the coefficients in $f_j(x)$.

By fixing the initial number of component densities of the mixture N we have for the beanplot data:

$$g(x|\Psi) = \sum_{j=1}^k p_j f(x|\theta_j) + \eta \quad (7.5)$$

where η is a specific associated error and $\Psi = (p_1 \dots p_k, \theta_1, \dots, \theta_k)'$ are the complete list of coefficients of the mixture model.

We fix both K and h and we derive the density model time series as a mixture of different distributions.

We retain the **coefficients** p_j representing the different components of the distributions occurring in the mixture.

In particular, the estimation method used is the maximum likelihood method, where it is necessary to consider some starting points (obtained by the beanplot analysis). A complete overview of the method used is in Du 2002 [239].

7.2. Coefficients Estimation: The Mixture Models Approach

So we can build t vectors of k coefficients, A_t , and substitute the beanplot time series (BTS) with them.

Therefore A_t is defined as:

$$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'. \quad (7.6)$$

We also consider a measure of the **goodness of fit** I_t (the Chi-squared statistic) representing the **quality of the representation (diagnostic phase)**.

We start from a simulated time series at time t in which we compute the beanplot observation b_t . We estimate the coefficients of the observation (table 7.2 and Algorithm 8). So we obtain:

Table 7.2: Coefficients estimation example

	pi	mu	sigma
1	0.50	2.94	2.02
2	0.50	5.00	1.02

And in particular, the coefficients are:

$$A_t = [0.5, 0.5]'. \quad (7.7)$$

We take into account the parameter sequences that could be statistically tested for checking the **structural changes** over time t . In particular for each coefficient for the associated time series we can consider a Chow test. We consider each coefficient in A_t , $p_{1,t}, p_{j,t}, \dots, p_{k,t}$.

We estimate each model of the coefficient time series in the sense:

$$p_{j,t} = \beta_0 + \sum_{q=1}^Q \beta_q \eta_q + \omega_j \quad (7.8)$$

where η_q is a dummy variable $(0, 1)$ that represents a specific period or an interval period of time, in which the null hypothesis of no structural change is tested. In the presence of structural change $\beta_q \neq 0$. At the same time ω_j is a residual. We return the dates of the structural changes for all the coefficients in A_t

Data: A Beanplot time series (BTS) $\{b_{Y_t}\} t = 1 \dots T$

Result: A vector A_t with the p as the proportions of the mixture components, given the bandwidth h

begin

 Choice of the I temporal interval

 Choice of the n points

 Choice of the h bandwidth to use

for $t \in T$ **do**

 | Estimating the coefficients of the mixture model for the
 | Beanplot B_t

end

 Are the internal models not fitting data adequately?

if *the internal model is not adequately fitted* **then**

 | change the number of coefficients p or the interval
 | temporal I

end

end

Algorithm 8: Beanplot internal modelling: coefficients estimation

7.2.1 Choosing the optimal interval temporal

It is necessary to consider as the optimal, the temporal interval related to the lowest error. In practice we compare different specifications over the time and we decide on the best one by considering the index of

goodness of fit over the time. We decide this by combining the different goodness of fit (using for example, a mean of the different goodness of fit) and we choose the interval temporal that minimizes this index. The algorithm is shown in algorithm 8 (page 220).

7.3 Beanplot Representations by their Descriptor Points

Here we can obtain and specifically represent the beanplot shape. In particular we assume that the original beanplot shape is constituted by two parts: a structured one (defined as a "model") plus a residual (in this sense we follow the approach to histogram approximation in Signoriello (2008)):

$$B_t = M_t + E_t(t = 1...T) \quad (7.9)$$

where B are the real data (the beanplot representations in the case) at temporal interval t , M represents the model, and E is the residual part. We need to represent the structured part M_t and minimize the residual part E_t .

We know that for every probability density function $\phi(x)$, for any probability density function the area is represented by:

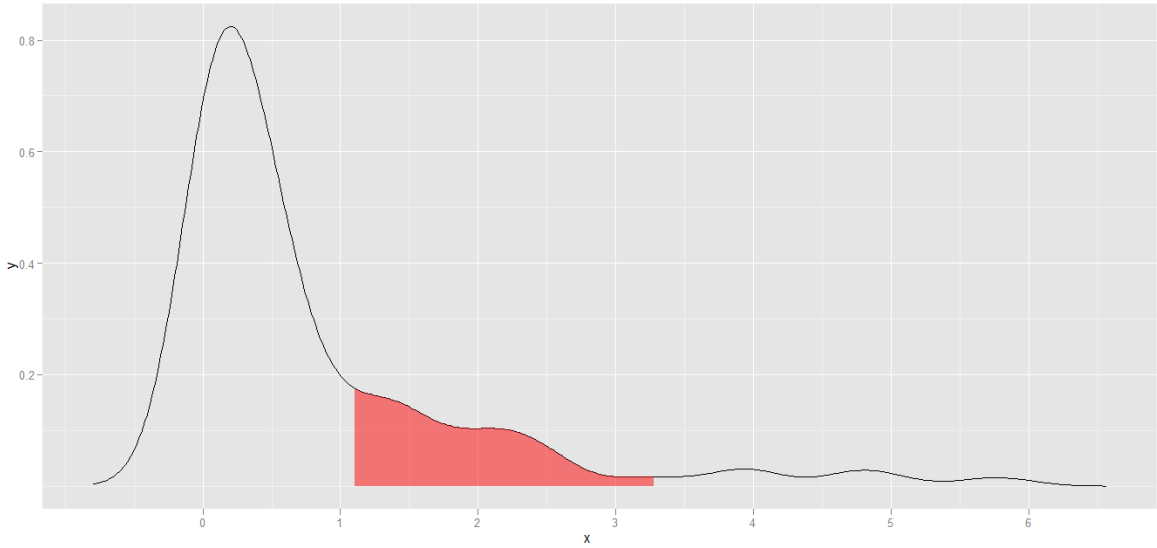
$$\int_{-\infty}^{\infty} \phi(x) dx = 1 \quad (7.10)$$

Or also:

$$\hat{f}_{h,t} = \frac{1}{nh} \sum_n^{i=1} K\left(\frac{x - x_i}{h}\right) = \int_{-\infty}^{\infty} \phi(x) dx = 1 \quad (7.11)$$

In that sense we can estimate, in the case of no autocorrelation between the observations, the probability that a specific random variable Z lies between z_1 and z_2 figure 7.5. So we have:

Figure 7.5: Kernel Density Estimation: computing the area between z_1 and z_2



Following Hyndman 1996 [378] and Samworth Wand 2010 [604] we can define the regions of highest density¹⁵:

Definition (Hyndman 1996) [378] Assume the density function $f(x)$ of a random variable X . The $100(1 - \alpha)\%$ HDR is the subset $R(f_\alpha)$ of the sample space of X such that:

$$R(f_\alpha) = (x : f(x) \geq f_\alpha) \quad (7.12)$$

f_α is the largest constant, such that

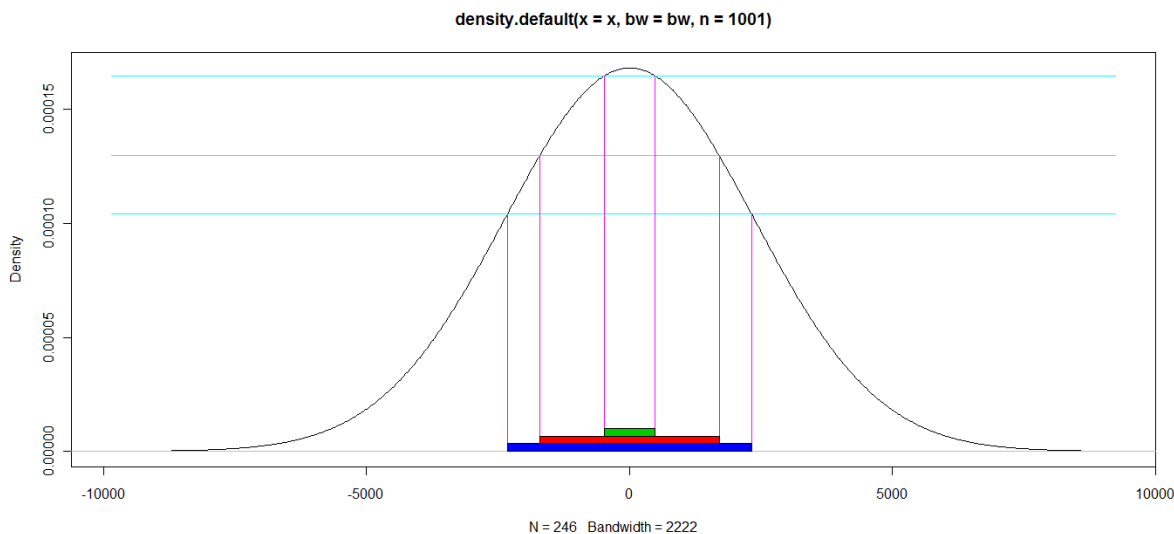
¹⁵see also Fadallah 2011 [263] for a short review

7.3. Beanplot Representations by their Descriptor Points

$$Pr(X \in R(f_\alpha)) \geq 1 - \alpha \quad (7.13)$$

It is possible to compute with the algorithm by Hyndman 1996 the regions of highest density (figure 7.6, table 7.3 and 7.4). In the calculation of the highest density, of great importance is the way in which the quantiles are computed (for the methods and the algorithm used see Hyndman 1996 [378]). At the same time, it is very important to compute the confidence intervals. Computing the quantiles for the density data could give a measurement of the risk, for example in financial data.

Figure 7.6: Highest Density Regions: BVSP Market (differenced series) year 2010



mode computed: 10.83034

As we know, an important difference between boxplot data and the density data is its ability to capture the multi-modality. So we can

Table 7.3: highest density regions (hdr)

	hdr.1	hdr.2
99%	-2311.09	2322.03
95%	-1698.08	1712.01
50%	-472.93	490.10

Table 7.4: falpha

	falpha
1%	0.00
5%	0.00
50%	0.00

compute also the HDR boxplots to observe the multimodality over time.

It is necessary to generalize the quantile computation from the original density data. In this sense we can have a specific quantification of the data over time. (see for example Sheather Marron 1990 [623]).

Definition 6. The Beanplot Time Series (BTS) descriptor attributes are realizations of the single features of the beanplot $\{b_{Y_t}\} t = 1 \dots T$ as coordinates X^C and Y^C . We refer to them as descriptor points, in which we measure the beanplot structure. Each beanplot can be represented by considering either the coordinates of n points for the X^C describing the location and the size (the support of the density) or the Y^C describing the shape (or the density trace).

So we obtain from the X^C :

$$X^C = [x_{1,t}, x_{j,t}, \dots, x_{k,t}]' \quad (7.14)$$

And from the Y^C :

$$Y^C = [y_{1,t}, y_{j,t}, \dots, x_{k,t}]' \quad (7.15)$$

To specifically represent the beanplot we choose firstly the number n of descriptor points to represent the initial density and then we obtain numerically the coordinates X^C and Y^C (Algorithm 9). The X^C represents the values in which the density Y^C is calculated.

In particular the Y^C value represents an estimate of the probability density at the point represented by X^C value. If we consider a higher number of points n in the procedure we obtain a more precise approximation of the original beanplot. In practice the choice of the number of points to describe the density corresponds to the choice of the number of bins in a histogram. However when we use a kernel estimation procedure we have higher flexibility than in the histogram case to separate the structure and to decide the optimal number of points. The choice of the descriptors n is a problem of finding the exact representation of the underlying phenomenon: are we interested only in a stylized image of the beanplot or do we need to represent all the features of the beanplots? In general are we correctly representing the underlying data? The different choices of the bandwidth and the kernel can improve the representation of the data, so when we obtain a satisfying data representation we can proceed to forecasting the process (algorithm 1).

The choice of h can be constrained to a number of descriptor points n to compute. A higher h (computed by the Sheather-Jones Method) can be an indicator of structural change in the internal model, therefore a coherent choice of the h needs to be made. By choosing a higher

Data: A Beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$

Result: A vector with n elements of X^C and Y^C coordinates
given the bandwidth h

begin

Choice of the n points to represent

Choice of the h bandwidth to use

for $t \in T$ **do**

 | Estimating the X^C

 | Estimating the Y^C

end

Is the internal model not fitting data adequately?

if *the internal model is not adequately fitted* **then**

 | change the number of descriptor points n or the
 | bandwidth h

end

end

Algorithm 9: Beanplot internal modelling: representation by descriptor points

7.3. Beanplot Representations by their Descriptor Points

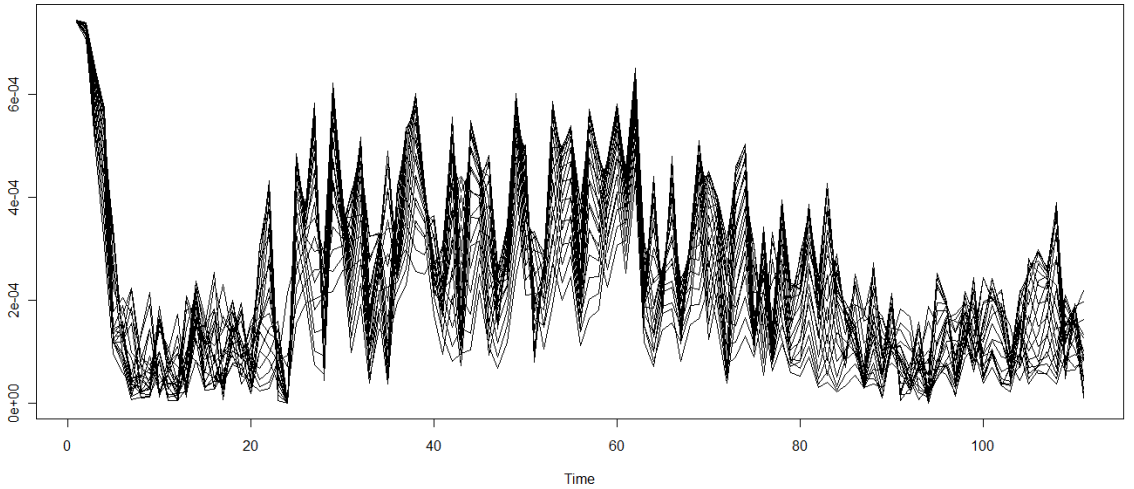
h and a higher n we can specifically take into account the complexity of the beanplot model.

7.3.1 Descriptor point interpretation: Some Experiments on Simulated and real datasets

By experimenting the representation on the coordinates, the important point we can note is that, differently from the first type of modellization, here we try to give an accurate description of the initial beanplot time series (BTS) (figure 7.7 and figure 7.8).

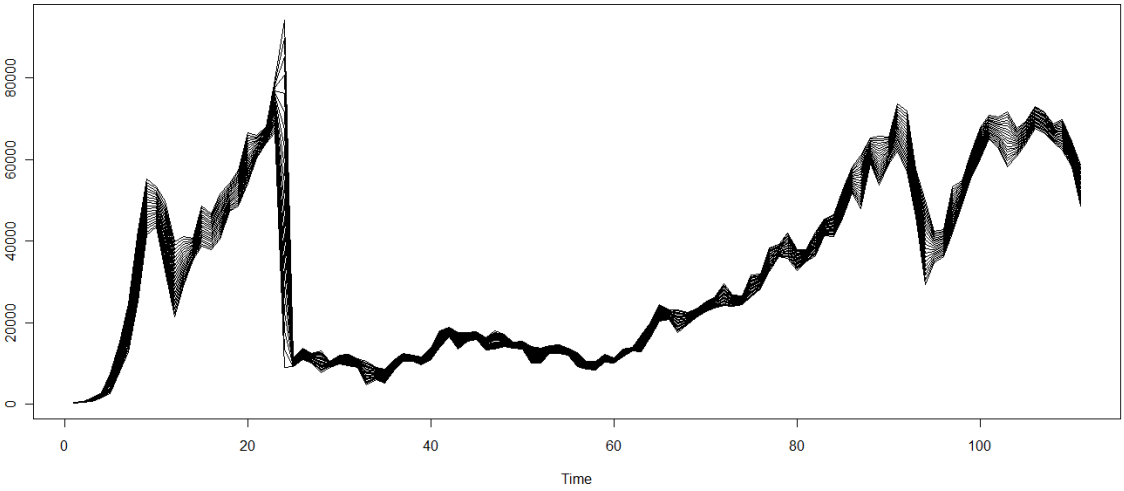
The representation on the X^C describes the location and dynamic size characteristics of the beanplot time series - BTS (the dynamic intra-period variability).

Figure 7.7: Bovespa Beanplot Time Series -BTS Y^C attribute time series of the descriptor points 1993-2011 ($n = 20$)



The Y^C represents the changes in the shape over time, so it is possible to detect structural changes. So we can observe a high volatility of the descriptor points due to changes in the density traces.

Figure 7.8: Bovespa Beanplot Time Series (BTS) X^C attribute time series of the descriptor points 1993-2011 ($n = 20$)



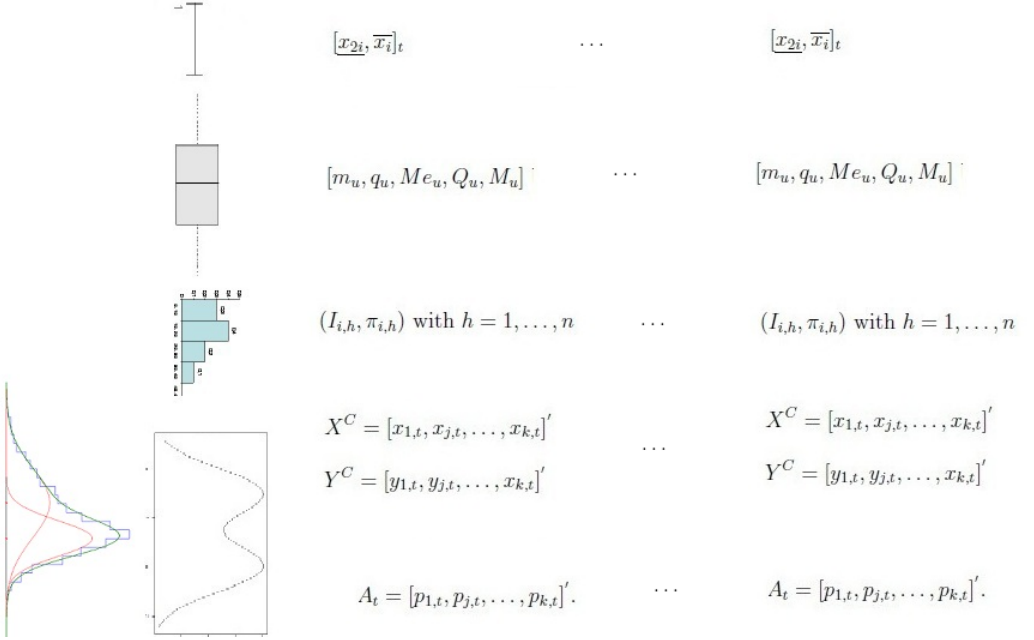
In practice by comparing the two representations it is possible to detect structural changes. It is interesting to note that by using the X^C representation it is possible to see the characteristics of the original time series.

Finally it is possible to compare the different representations by their descriptors in the figure 7.9:

Initial time series can be represented as intervals (ITS), boxplots (BoTS), histograms (HTS) descriptors or beanplots (BTS) by estimating the coefficients A_t or considering the coordinate descriptors X^C and Y^C . The interval time series (ITS) is represented by its upper

7.3. Beanplot Representations by their Descriptor Points

Figure 7.9: Comparing descriptors amongst all the Representation Time Series (ITS, BoTS, HTS, BTS and models)



and lower bounds. The boxplots time series (BoTS) are represented by the quantiles, etc. The estimation by coefficients in the beanplot case tend to extract the relevant information where the information is the most synthetic amongst the other representations. Intervals consider for the interval period only the upper and the lower bound.

7.4 Data Tables considering Density representations

It is possible, now, to define what the coefficients and the descriptor points contribute in the generation of a specific data matrix that could be statistically analyzed.

In this respect, by starting from the classical time series we obtain the symbolic time series and the symbolic time series data table ([211] and figure 7.10). It is important to note that handling different outliers can be very important because we can obtain different results by taking into account different symbolic data. In this specific case the symbolic data analysis came directly from the data. Through choosing the best statistical descriptor we analyse these series using symbolic data analysis. See also Gettler-Summa M. Frédérick V. (2010) [299]. The objective of this thesis is to handle symbolic data which can be densities.

Following Signoriello 2008 [630], the function considered (or the statistical model in the wide sense) is substituted by the estimated coefficients and the descriptor points, so we obtain a final data table which is useful for the statistical analysis of the intra-variation between different models using the appropriate techniques.

In the data matrix of the estimated coefficients the data can be considered as *units* \times *variables* \times *number of coefficients*.

Each different function or model needs to be accompanied by the goodness of fit, or its ability to fit the original data. So in each cell we need to have k coefficients and an index (I) of goodness of fit. From the models and their goodness of fit it is possible to obtain a summary of the intra-period variation and the presence of models that are not well approximated

	Variable 1	Variable 2	...	Variable p
Observation 1	$b_{111}, b_{112}, \dots, b_{11k}, I_{11}$	$b_{121}, b_{122}, \dots, b_{12k}, I_{12}$...	$b_{1p1}, b_{1p2}, \dots, b_{1pk}, I_{1p}$
Observation 2	$b_{211}, b_{212}, \dots, b_{21k}, I_{21}$	$b_{221}, b_{222}, \dots, b_{22k}, I_{22}$...	$b_{2p1}, b_{2p2}, \dots, b_{2pk}, I_{2p}$
:	:	:	:	:
Observation n	$b_{n11}, b_{n12}, \dots, b_{n1k}, I_{n1}$	$b_{n21}, b_{n22}, \dots, b_{n2k}, I_{n2}$...	$b_{np1}, b_{np2}, \dots, b_{npk}, I_{np}$

Figure 7.10: Table of the parameters of the data models (Signoriello 2008 [630])

7.5 From Internal to External Modelling

Forecasting symbolic data is a growing field related to real problems fundamental to modern financial systems. We have proposed a method related not to scalar time series but to a new type of complex data such as density or beanplot data. Various families of questions are already open: the parameterization through densities by using different methods like the B-Spline, the forecasting by using alternative methods such as K-Nearest Neighbour by Arroyo, Gonz  les-Rivera, Mat  , Munoz San Roque (2010) [39] and the construction of dynamic factors related not to with scalar time series series but to beanplots (in order

to improve forecasting by using different beanplot time series BTS).

We need at this point to compute a way to tell if the forecasts obtained by the different models are good or not. So we build indices to study the level of the adequacy of the forecasts obtained by the different models.

7.5.1 Detecting Internal Models as Outliers

It is possible to use a clustering procedure to identify the outliers. In practice we use a distance (for example, the euclidean distance on the coordinates and the model distance on the coefficients of the mixture) to identify the different models that could be considered outliers.

In this sense we can conclude that the elimination of the noise is not successful and that we need to eliminate these extreme observations.

So we can proceed by either not considering the extreme observations, or by imputating them. In fact, we need to analyze the external models and either consider or not these extreme values (identified here as outliers).

A second strategy starts from the estimated coefficients and the descriptor points so as to identify from the trajectories the outliers (using a method that identifies outliers from the time series¹⁶). In this respect we identify outliers directly from external models.

7.6 Internal Models Simulation

From the Internal Models we can simulate the same internal models. This is useful for various reasons. In general we start from the primary data and we try to simulate n times the data generating process of the complex data through obtaining the data characteristics than can

¹⁶For example see Chang, Tiao and Chen (1988) [130] and Riani 2004 [582]

7.6. Internal Models Simulation

mimic the initial data. If the difference is small that means that the Noise part is not relevant. The structural part in this sense could be detected as a representation in the internal model as a density data. The aim of the simulation exercise is to simulate different scenarios and understand if the most important and relevant features of the data are captured by the models.

In particular the important question is: are the models good for identifying important data features of the reality, can we understand the structural part of the initial data and separate it from the noise?

Summary Results: internal modelling
Two approaches are used in the Internal Modelling phase.
The first approach assumes the data to be a mixture, and so the coefficients representing the components are taken into consideration. In this way, we extract the structural information by the Beanplots.
A TSFA model is used to synthesize the trajectories and so obtain the latent factor related to the shocks changing the Beanplot structures over time.
The second approach represent Beanplots as a whole and uses coordinates to represent simultaneously the Beanplot attributes.
The attribute time series or the trajectories of the descriptor points for the second representation are considered.
In both cases, coefficients and descriptor points substitute the original data.

Chapter 8

Beanplots Time Series Forecasting

Scalar Time Series (STS) Analysis and Forecasting¹ is the analysis of the characteristics of a time series and the application of a statistical model to forecast its future values.

Very relevant research developments have occurred in this field and important results have been obtained since the 1970's (De Gooijer and Hyndman (2006) [180]). Forecasting of time series data² is a difficult task when the data are very numerous, with complex structure³ for example, when there are high volatility and structural changes (for the

¹Hamilton 1994 [333] Lutkepohl 2005 [469], Battaglia 2007 [66], Durbin Koopman (2001) [244] and Elliott and Timmermann (2008) [247]. An important review work on Forecasting is De Gooijer Hyndman (2006) [180]

²Box Jenkins 1976 [99] Wallis 1989 [767]

³The financial markets for example are considered a complex system (in this sense: Marschinski and Matassini 2001 [488], Mauboussin 2002 [496], Sornette 2004 [636] and Tjung, Kwon, Tseng, Bradley-Geist (2010) [661]. Of relevance is the difference between signal and noise provided by the same operators that impacts on price, see for example the experiment in Cipriani and Guarino 2005 [144]

problems and the approaches in forecasting financial data see Deistler Zinner 2007 [187]). This is the case of high frequency data or, in general, of financial data, where we cannot clearly visualize the single data⁴ and where the necessity of an aggregation arises (see Chapter 1 and Chapter 2). A first proposal is that of using in addition to classical scalar time series some "density based" time series (see Chapter 5 for the characteristics of the density data) and in particular beanplot time series (BTS). In this chapter we deal with the specific problem of forecasting the beanplot time series (BTS). So we propose an approach based firstly on internal models (as we have seen in Chapter 7) of the beanplot time series (BTS) and successively on the choice of the best forecasting method with respect to our data. In particular we propose a strategy to use combination forecast methods⁵ in order to improve the statistical performance of our forecasts.

8.1 Density Forecasting and Density Data

There are important differences between density forecasting⁶ and the forecasting of density data. Density forecasting can be defined as "a density forecast of the realization of a random variable at some future time is an estimate of the probability distribution of the possible future values of that variable" (see Tay Wallis 2000 [653]). In the analysis of the density data we consider the densities for each temporal interval as data, and we build external models on them.

In the analysis of complex time series, such as high frequency data

⁴Zivot (2005) [722]

⁵See for a discussion of these methods Timmermann 2006 [658], Clemen (1989) [145] Armstrong 2001 [29] who review the method. For a formal explanation on why forecast combinations of single forecasts tend to outperform sophisticated models see Smith and Wallis 2009 [635]

⁶Timmermann 2000 [657] Bao Lee and Saltoğlu 2007 [59] Diebold Gunther 1997 [220]

or financial data with a particular data structure, it is not always possible to have an effective visualization and reliable forecasting. In forecasting, the problem of the visualization and the choice of the model to use are specifically linked, in fact, in identifying the adequate one it can be important to detect outliers. The problem of the visualization of complex time series has been particularly explored in a paper proposed by Drago and Scepi in 2009 [237] where we underlined the necessity of searching and analysing an aggregate behaviour for complex data and we have explored aggregated beanplot time series (BTS) to observe their characteristics. In particular we have shown the proprieties of beanplot time series (BTS) by comparing different types of aggregated time series and by illustrating their statistical performances in financial data interpretation. In this chapter we propose an approach for forecasting beanplot time series (BTS). In particular, after a short introduction on the definition and the main characteristics of beanplot time series (BTS), we deal with the problem of finding an appropriate internal model to define an external model with the aim of forecasting beanplot time series (BTS). We propose our forecasting approach in Sections 9.3 and 9.4 in a summarized form while we illustrate it in greater detail in the applicative section. We show that the advantage of our approach is to forecast not only the average or the location of the data but at the same time to predict the size and the shape of the data (the entire structure). In the specific case of financial and high frequency data we are trying to genuinely predict the associated risk or the volatility that can occur over time. Size and shape can represent the internal variation over the interval temporal, so, in this sense there is a specific link between the data collection and quantitative methods used: we try to forecast the market instability or the internal variation in the data. For density forecasting see Hyndman and Fan (2008) [387]. Tay Wallis (2000) [653]

8.2 From Internal Modelling to Forecasting

The Beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$ is an ordered sequence of beanplots or densities over time (see Drago Scepi 2009 [237]). The time series values can be viewed as realizations of an X beanplot variable in the temporal space T , where t represents the single time interval. For forecasting purposes the choice of the length of the single time interval t (day, month, year) depends on the specific data features and objectives the analyst wants to study. In practice, detecting the adequate model means finding the outliers in the series before moving on to the analysis, there is also the need to consider structural changes in the model. For this reason it is very important to consider visualization and clustering methods (in particular with the objective to detect outliers⁷ before proceeding to the analysis).

Various approaches could be considered:

1. Considering trajectories from the coefficients of the mixtures (Chapter 8.3)
2. Considering the trajectories from the beanplot descriptions or attribute time series (Chapter 8.4)
3. K-Nearest Neighbour (Chapter 8.5)
4. Forecast Combinations (Chapter 8.6)
5. Hybrid approaches (Chapter 8.6)
6. Uses of GA (Chapter 8.6)
7. Forecasting using the Search Algorithm (Chapter 8.7)

⁷In scalar data: see for example Cherednichenko 2005 [132]

8.3 External Modelling (I): TSFA from model coefficient approach

We start from an internal model described, like the coefficients of the extracted mixtures, in Chapter 4. In this sense we are considering the sequence of the coefficients over the time. The beanplots are entirely substituted by these coefficients sequences.

We can forecast the next beanplot at $t + 1$ by considering a time series forecasting method. Starting from the Time Series Factor Analysis (in particular Meijer Gilbert 2005 [500]), in practice, we can define each vector of coefficients as a combination of time factors:

$$p_t = \alpha + V\xi_t + \epsilon_t \quad (8.1)$$

with α as a vector of intercepts, V as a $J \times M$ matrix of factor loadings and ϵ_t as a vector of random errors. So, by considering K observed $p_{j,t}$ with $j = 1..J$ and $t = 1..T$, in which we are searching the M factors as unobserved processes $\xi_{m,t}$ with $m = 1..M$ and $t = 1..T$. Therefore, for each time series we can obtain a number $q \leq k$ number of factors. The loadings L are estimated by FA estimators (such as ML), by using the sample covariance of the error (Wansbeek Meijer 2005 [689])

To measure the factor, we use:

$$\xi_t = (V^t \Psi_{-1}^t V)^{-1} V^t \Psi^{-1} V^t (p_t - \alpha_t) \quad (8.2)$$

In which $\psi_t = Cov(\epsilon_t)$. In particular the loadings, in the FA estimators (say, ML), can be estimated by using the sample covariance of the error.

The usual assumption made in TSFA is that $\alpha_t = 0$, so the Bartlett predictor becomes:

$$\xi_t = (V^t \Psi_{-1}^t V)^{-1} V^t \Psi^{-1} V^t (p_t). \quad (8.3)$$

In this way, it is possible to compute the factor time series for the factors related to the location and the size and another factor representing the shock response or the short run dynamics. We can use different forecasting methods for forecasting the different time factors (starting from an ARMA model). In particular, for forecasting factor time series using the TSFA see Muñoz , Corchero and Javier Heredia (2009) [523]:

$$\hat{\xi}_t = c + w_t + \sum_{s=1}^r \varphi_s \xi_{t-s} + \sum_{s=1}^q \theta_s w_{t-s}. \quad (8.4)$$

Starting from the prediction of the factors, from eq. (4) we can compute the initial coefficients by giving the value of the $\hat{\xi}_t$. So we have:

$$\hat{p}_t = \alpha + V \hat{\xi}_t + \epsilon_t \quad (8.5)$$

From the predictions of the initial coefficients we can fit the predicted beanplots. Assuming the number of initial groups of mixtures $N_{j=1\dots k}$ we have for the beanplot:

$$\hat{B}_t = \sum_{j=1}^k \hat{p}_{j,t} N_{j,t} + \eta_t \quad (8.6)$$

Where η_t is a specific associated error.

From the predictions of the coefficients we can fit the observed beanplots \hat{B}_t . Poor results in in-sample forecasting can occur in a revision of the initial parameters considered (the kernel K used and the bandwidth h for example). At the same time, poor forecasting results can be useful in order to identify structural changes, outliers or abnormal influent observations.

8.3. External Modelling (I): TSFA from model coefficient approach

We use the combination of the forecasts to improve the forecasts in the presence of model parameter drifts and uncertainty in the identification procedure

$$F_t^{CM} = \gamma_1(t)f_t^1 + \gamma_2(t)f_t^2 + \cdots + \gamma_m(t)f_t^m \quad (8.7)$$

In the combination of the forecasts we use $f_1 \dots f_m$ forecasting techniques (i.e Exponential Smoothing⁸, Spline⁹, Theta methods¹⁰). Various strategies of the weight estimation of the combination model $\gamma_1 \dots \gamma_m$ are used. By considering explicitly $f^1, f^2 \dots f^m$ as different forecasts from competing external models we have as external model the forecast combination F_t . In this case $\gamma_1 \dots \gamma_n$ are different weights. In particular the weighted combination¹¹ of the models allows one to improve forecasting in a context of parameter drift and structural change.

Poor results in in-sample forecasting can occur in a revision of the initial parameters considered (the kernel K used and the bandwidth h for example), in order to identify outliers or abnormal influent observations. A search algorithm is used in the validation phases to improve the predictions by selecting the best relevant information as forecasting interval.

⁸Hyndman, Koehler, Snyder and Grose (2002)[384] Hyndman, Koehler, Ord and Snyder (2008) [385] Hyndman, Akram, and Archibald (2008) [380]

⁹ Hyndman, King, Pitrun and Billah (2005) [388]

¹⁰ Assimakopoulos and Nikolopoulos (2000) [45], Hyndman and Billah (2003) [381], Makridakis Wheelwright Hyndman (2008) [480]

¹¹Winkler Makridakis 1983 [703] Bates Granger 1969 [64] Newbold Granger 1974 [534]

8.3.1 Detecting Structural Changes

Coefficient sequences could be statistically tested for checking the **structural changes** over time t . We use a different Chow test¹² in every associated time series for each coefficient.

We consider each coefficient in $A_t, p_{1,t}, p_{j,t}, \dots, p_{k,t}$. We estimate each model of estimated coefficients time series in this sense:

$$p_{j,t} = \beta_0 + \sum_{q=1}^Q \beta_q \delta_q + \omega_j \quad (8.8)$$

where δ_q is a dummy variable $(0, 1)$ that is representing a specific period or an interval period of time, in which the null hypothesis of no structural change is tested. In the presence of structural change $\beta_q \neq 0$. At the same time ω_j is a residual.

We return the dates of the structural changes for all the coefficients to A_t

8.3.2 Examples on real data: Forecasting World Market Indices

This application has twofold objectives. The application is divided into two parts: We consider firstly the time series of the Dow Jones Index for the period 1990-2010 and we compare three different approaches in forecasting: scalar data, beanplots and interval valued time series. Secondly, we study the mechanisms of the financial contagion between the financial systems in the crisis by trying to forecast other world indexes.

We consider the case of various markets and the beanplot time series (BTS) related to the closing price. In practice, we examine various market indexes in which we consider the high, low, close, and open

¹²Chow 1960 [140] Harvey 1990 [343] Chu Stinchcombe White 1996 [142]

8.3. External Modelling (I): TSFA from model coefficient approach

price for various periods (depending on the data availability of each single market in Yahoo). For each dataset we consider the closing price and we obtain the beanplot time series (BTS) for the single market as well.

Then we model the beanplot by considering the beanplot as a mixture, and we obtain the sequence of the estimated coefficients over time. At this point we can consider a factorial strategy (Time Series Factor Analysis or TSFA) so as to obtain the factor time series representing the single market.

Then we can use a forecasting method, for example an ARIMA, to predict the market dynamics over time as represented in the factor considered (figure 8.1, figure 7.1, figure 7.1, figure 8.4, figure 8.5, figure 8.6)

It is interesting to compare the results with interval data for the case of the Dow Jones Index (DJI):

1. The scalar data does not allow us to forecast the **internal variation** in the year (useful for risk management purposes).
2. Interval time series (ITS) does not permit us to observe in the period the **intra period structural changes**. The beanplot bumps allow us to observe the internal variation also in this sense.
3. The **intra period seasonality** aspect is not represented by aggregating the time series (using a mean) or in interval time series (ITS). The different mixtures can be interpreted in some cases as structures related to intra period seasonality.

BEANPLOTS TIME SERIES FORECASTING

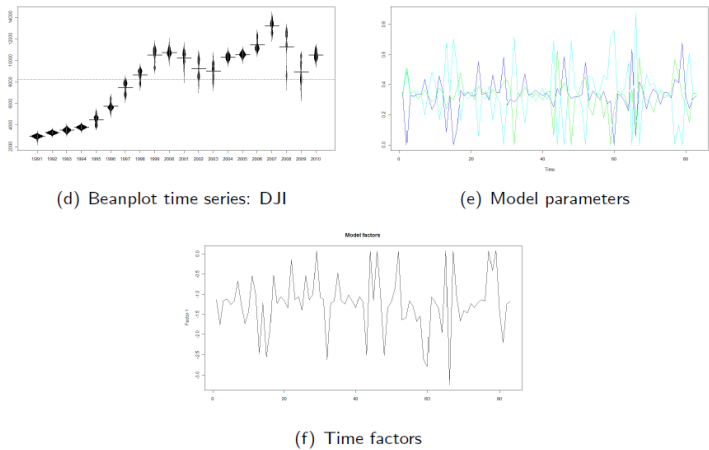


Figure 8.1: DJI - Dow Jones Index

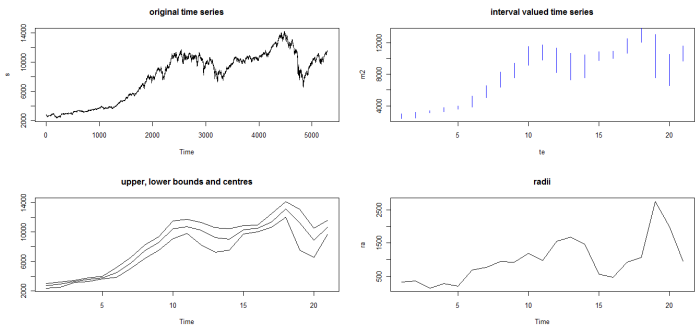


Figure 8.2: DJI - Dow Jones Index

8.3. External Modelling (I): TSFA from model coefficient approach

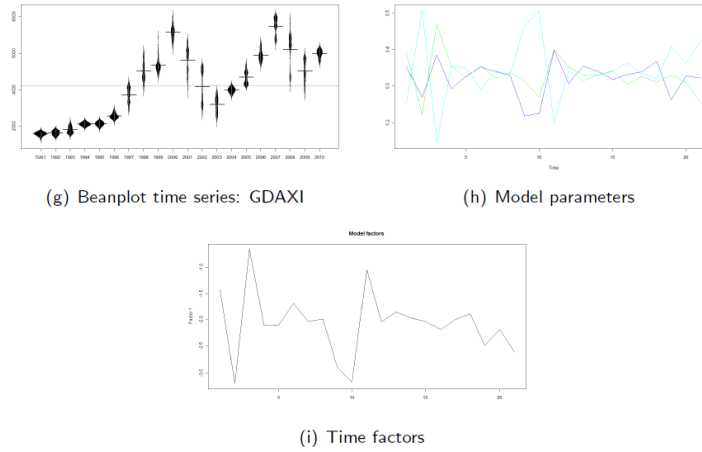


Figure 8.3: GDAXI - German Dax Index

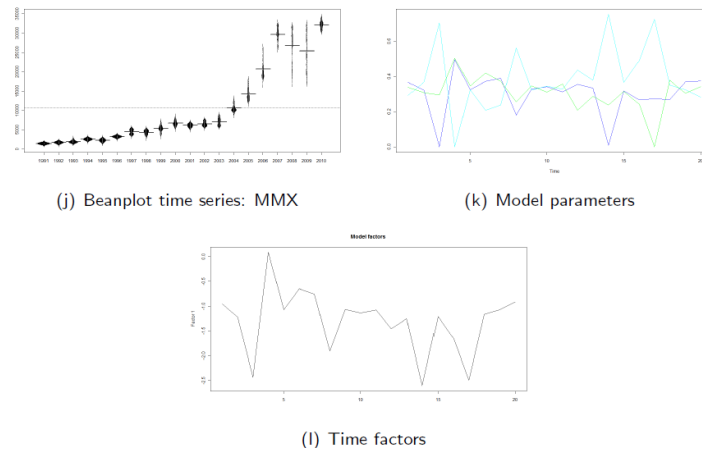


Figure 8.4: MMX - Major Market Index Mexico

BEANPLOTS TIME SERIES FORECASTING

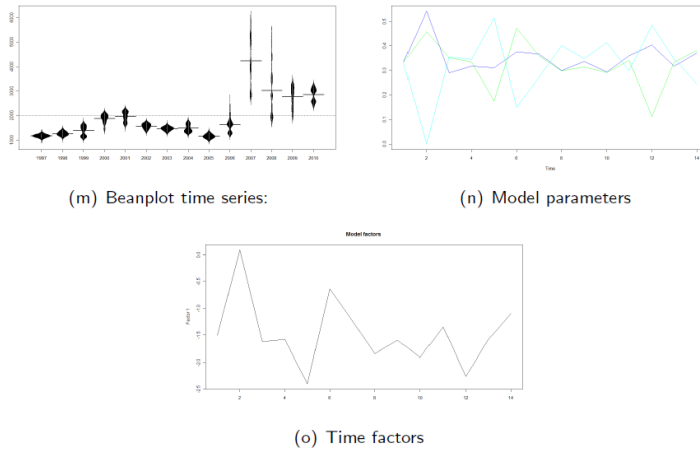


Figure 8.5: SSEC - China Stock Market Index

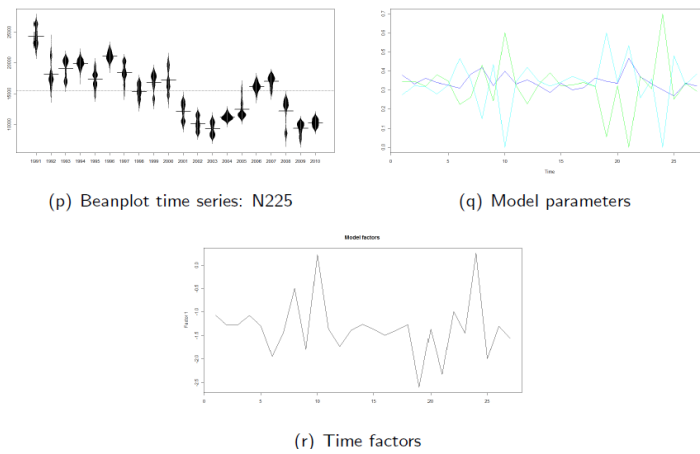


Figure 8.6: N225 - Nikkei Index Japan

8.4 External Modelling (II): Attribute Time Series Approach from Coordinates

The second forecasting method is related to the approach of the coordinates. We need to define an internal model before forecasting. The idea is to represent each beanplot by using specific attributes of each beanplot and by considering their realization over the time. So we define a Beanplot Attribute Time Series as a realization of a single beanplot $\{b_{Y_t}\} t = 1...T$ descriptor over the time. To represent the beanplot time series (BTS) it is possible to consider the coordinates X^C and Y^C as descriptors of the beanplot. We refer to them as descriptor points because they measure the beanplot structure. In particular, the X^C time series clearly show the location and the size of the beanplot over the time while the Y^C represents well the shape over the time. To specifically represent the beanplot we choose firstly the number n of descriptor points and then we obtain numerically the coordinates X^C and Y^C . If we consider a high number of points n in the procedure we obtain a more precise approximation of the beanplot, here it is necessary to note that to have a density we have to consider at least 4 points (but not in every situation: we could be interested in forecasting the entire density trace).

The choice of the descriptors n is a problem of the exact modelization of the underlying phenomenon: are we interested only in a stylized image of the beanplot or do we need to represent all the features of the beanplots?

In our approach we consider the values of X^C and Y^C corresponding to the 25th, 50th and 75th percentile. The procedure is coherent with existent literature (Arroyo 2009 [32]) in which, from the initial interval data, the author obtains the attribute time series for the upper and the lower bound. By this type of descriptor points we are able to detect the evolution over the time of the beanplots, for the location,

the size (both the X^C) and the shape (mainly the Y^C).

8.4.1 Analysis of the Attribute Time Series Approaches

Our approach can be synthetized in the following steps. The first step is the analysis of the structure of the sequence of the attributes. In this sense we have to test the structure of each attribute time series. As we know, they represent the beanplot dynamics over the time, so we can use a specific method to forecast the attribute time series to obtain the prediction at time $t + 1$, $t + 2$ and so on. Initially we can detect the underlying structure of the data by decomposing the attribute time series. This decomposition can be very useful both to understand the general dynamics of the series over time (represented for example by the trend) and to exploit some interesting patterns.

Successively we must decide which model has to be used for the forecasting approach. We can use univariate or multivariate methods. In the first case we are assuming there is no specific relationship between the attribute time series, in the second case we are assuming that a relationship exists (in practice we analyze attribute time series using time series analysis). So, it is firstly important to test the stationarity of the attribute time series and successively to define the possible cointegration between the series. Another important point is to verify the autocorrelation of the attribute time series and the possible structural changes. Only at the end of this first complex analysis can we decide which model should be used for forecasting our attribute time series. After the identification of the forecasting models, it is necessary to estimate the different models for obtaining the forecasts and, finally, to evaluate the reliability of our forecasts. The diagnostic procedure is very important because we can critically evaluate and respecify the models. At the same time it is very useful

to consider the performances of the different forecasting models by considering some evaluation indexes. Using more than one forecast may be necessary to obtain better predictions (see Timmermann 2006 [658]).

So in general, an approach of forecast combinations may be necessary for many reasons. First of all there can be some structural changes and it may be necessary to take into account more than one forecasting model. In this sense, the use of different models can reduce the risk associated to choosing one single model. Secondly, there can be some uncertainty in the model to use, so it is very important to define a combination of forecast strategies to define the best model. Various approaches can be followed in setting the combination method. One strategy is to consider many different forecasts from several models and to use an average of them. Here, a relevant problem is the choice of the different weights to apply on the different forecasting methods. Assigning higher weights to methods that permit the improvement of the accuracy of the forecasts (by minimizing errors) can be a solution. Finally, a relevant point could be the choice of the optimal interval for the forecasting process.

8.4.2 Attribute Time Series Forecasting Models

Here we start to test the attribute time series. In particular we perform the usual tests on nonstationarity for either the Y^C and the X^C attribute time series on different subperiods. At the same time we decompose the attribute time series to find some patterns that could be useful in the modelling phase. It is very interesting to note that the beanplots quickly change their features over the time, in particular the shape, represented in the Y^C attribute time series. So, for the modelling purposes of the attribute time series we consider the longest interval, and we test nonstationarity. We obtain the result that both the X^C and the Y^C attribute time series are nonstationary.

Furthermore we perform the Cointegration Analysis, considering the attribute time series on the X^C and those on the Y^C . The final aim of this part is to investigate if in the case of the X^C or in the Y^C we can use a Vector Error Correction Model in forecasting (see also Arroyo 2009 [32] for an approach based on interval time series - ITS). To take in to specific account the structure of the single attribute time series, we identify various different models as starting forecasting models. At this point we estimate the different models also by considering the different subperiods for robustness checking. The different forecasting models are evaluated in terms of different indexes of Forecasts Accuracy¹³(see Hyndman Koehler 2006 [386]). At the same time if we can predict the density data at the end we need to check if the integral of the area under the prediction is equal to 1. In this case the prediction of a density could be considered.

8.4.3 Identification and External Modelling Strategy

So we can summarize in this way the entire forecasting procedure using the coordinate descriptor points in this way (Algorithm 10 and Algorithm 11):

1. Start to consider the n attribute time series of the descriptors (*i.e.* $x_{1,t}, x_{2,t}, x_{3,t}, y_{1,t}, y_{2,t}, y_{3,t}$) of the beanplot time series (BTS) b_{Y_t} for $t = 1, \dots, T$
2. The attribute time series represent the external models (the dynamics over the time $t = 1, \dots, T$) where each beanplot can be considered as the internal model at time t

¹³See also Armstrong 2006 [28]

8.4. External Modelling (II): Attribute Time Series Approach from Coordinates

Data: A set of a attribute time series for the beanplot time series (BTS) $\{b_{Y_t}\} t = 1 \dots T$ each one representing a different feature related to X^C or Y^C , a set of r forecasting methods in R , a set of combination schemes cs in CS

Result: A set of Forecasts f_k with $k = 1 \dots n$

```

begin
  for  $a \in A$  do
    Stationarity and Nonstationarity tests on  $a$ 
    Statistical decomposition of the time series  $a$ 
    Is it possible to forecast  $a$  using the causal forces?
    if it is possible then
      | forecast  $f_1$ 
    end

    Autocorrelation tests on  $a$ 
    Is it possible to forecast using the autocorrelation
    structure?
    if it is possible then
      | forecast the series and obtain  $f_2$ 
    end

    Analyzing the relationships between the attributes
    for  $r \in R$  do
      Forecasting using different methods  $r_n$  with
       $n = 1 \dots k$ 
      Obtaining different  $f_n$  with  $n = 1 \dots k$  forecasts
      Compute the adequacy of the single forecasts  $f$ 
    end
  end
end

```

Algorithm 10: Identification, External Modelling and Combination
(1)

Data: A set of a attribute time series for the beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$ each one representing a different feature related to X^C or X^Y , a set of r forecasting methods in R , a set of combination schemes cs in CS

Result: A set of Forecasts f_k with $k = 1...n$

begin

Is it possible to improve the forecasts using combination schemes cs ?

if *it is possible* **then**

| Use a combination scheme cs

end

Is it possible to improve the forecasts using weights?

if *it is possible* **then**

| Use combination weights F_2

end

if *it is possible* **then**

| Seek weights w

| Use combination weights

end

Is it possible to improve the forecasts using a search algorithm?

if *it is possible* **then**

| Search and apply the best set in the forecasting

end

end

Algorithm 11: Identification, External Modelling and Combination
(2)

3. Start to consider the n attribute time series of the descriptors of the beanplot. They represent the beanplot dynamics over the time
4. Checking for the stationarity and the autocorrelation. Detecting the features of the dynamics (trends, cycles, seasonality). Analyzing the relationships between the attributes
5. Forecasting using different methods
6. Beanplot Forecasts combinations schemes
7. Search Algorithm to improve forecasts
8. Considering as Beanplot description the forecasts obtained from the Forecasting Method.
9. Checking if the area of the predicted descriptor points can be a density (the integral is equal to 1).

8.4.4 Examples on real data: Forecasting the Beanplot Time Series (BTS) related to the Dow Jones Market

The application is related on the Dow Jones market dataset from the year 1928 to the 2010, and the purpose is to forecast the beanplot time series (BTS) for the specific period 2009-2010. In particular we consider in the models the data from 1998 to the 2008. The entire period in the database (1928-2010) is considered only for exploratory purposes. The specific aim in the forecasting process is to predict the instability in the market over the time. Methods used in the forecasting models are Smoothing Splines¹⁴ and Automatic Arima¹⁵ primarily,

¹⁴ Hyndman, King, Pitrun and Billah (2005) [388]

¹⁵ Hyndman and Khandakar 2008 [382]

and also VAR¹⁶ (Vector Autoregressive Models) and VECM¹⁷ (Vector Error Correction Models) in the case of the Y^C attribute time series in some specific periods to analyse the robustness of the models (we work on different sub-periods to analyse the existence of structural changes in the period). In a second forecasting model we use a combination approach, by combining different forecasts obtained by different methods. Lastly, we compare all the results, using as a benchmark: the Naive model (that is, representing the prediction obtained by considering the last observation).

So as a first step in the procedure we visualize the entire beanplot time series (BTS) by considering the yearly temporal interval (where this interval is the most appropriate in the data exploration). We can clearly visualize the existence of the two crises, in particular the crisis in 2008, where we can observe the very peculiar shape of the beanplot, in that it is strongly stretched out. The rise and the fall of the New Economy in 2001 is visible as well in the data. At the same time it is possible to detect the trends and the cycles of the beanplot time series (BTS) where we can also understand the structural change occurring over the time (for example in 2008).

We compute the set of X^C attribute time series (year 1996-2010), starting from the representation by coordinates. Here we can consider a different interval temporal from the interval used in the data visualization (that for considering the statistical features of the beanplots). At the end of the internal modelling process we obtain 6 attribute time series (3 for the X^C and 3 for the Y^C) and around 60 observations (years 1996-2010 approximately).

The three X^C attribute time series are related to the 25th, 50th and the 75th percentile where each Y^C is associated to the X^C . We call these intervals "extreme risk intervals", with minima or lower risk,

¹⁶ Lutkepohl H. (2005) [469] and Hamilton (1994) [333]

¹⁷Lutkepohl H. (2005) [469]

median risk and maxima or upper risk. These intervals are directly related to the beanplot structure. The attribute time series for the X^C show the long run dynamics of the beanplots and also the impact of the financial crisis. By considering the Y^C attribute time series (year 1996-2010), we need to remember that we are examining a different interval temporal (not the yearly interval temporal) but two months (around 40 observations in a beanplot data). In this representation we can observe the complexity of the initial series, observed in particular in the changes time by time of the beanplot shapes (represented by the Y^C). These behaviors are related to the short run dynamics of the series. Now we can begin to test the stationarity for the attribute time series.

8.5 The K-Nearest Neighbour method

Here we can consider a method based on the K-Nearest Neighbour algorithm. In particular we depart from a data matrix of the single estimated coefficients or the descriptor points, and we forecast the different beanplots in order to maximize the similarity of the periods over time and consider a median of the values in time.

Following Arroyo 2008 [32] and Arroyo, González-Rivera, Maté (2010) [41] it is possible to use the K-Nearest Neighbour to predict the minima and the maxima of the descriptor points of the beanplot. In practice we follow two separate steps:

1. We search for the n interval sequences that could be defined as closest to the current one.
2. The sequence extracted is necessary to consider this sequence as characterized by the last d intervals (the previous one).

3. Clearly it is necessary to consider a distance to measure the dissimilarity of interval sequences regarding appropriate distances (see Arroyo 2008 [32])
4. We consider the mean of the subsequent interval of the n closest sequences to obtain the forecast.
5. Arroyo 2008 [32] proposes the use in the mean computation of the interval arithmetic (see in that sense Moore 1966 [513], Gioia and Lauro 2005 [310] and Gioia 2001 [307])
6. Different weighting schemes can be applied

At the same time it is possible to consider the forecasts obtained as a first forecast in a more complex combination of forecasts. So it is possible to consider the predictions obtained using the K-Nearest Neighbour as one of the weighted components in the forecast combinations (see in the context of interval data Salish and Rodrigues 2010 [603]).

At the same time it is possible to use an approach on the entire beanplot considering each different attribute time series separately (see Yakowitz 1987 [709] for the approach of K-Nearest Neighbour in time series analysis).

8.6 The Forecasts Combination Approach

Furthermore we use a different approach based on the Forecasting Combinations. In particular, we use combinations for two reasons: we can take into account precisely the uncertainty in choosing a specific model (in fact sometimes the choice of a specification in a single forecasting model can be very hard) and we can also take into account the specific structural changes that could be captured (or we try to

capture) by considering combinations of forecasts. So we consider predictions obtained by using various methods: Smoothing Splines, Auto Arima, the mean of the period, the Theta method and the Exponential Smoothing. See table 1 to compare the different MAPE for the X^C .

We do not use any special weighting structure but we consider only the average between results obtained from the different methods. Regarding the results, the forecasting performances outperform other single models based on the single forecasting model chosen. Here we are considering an interval of prediction of 5 periods ahead (for this reason we do not apply this method to the Y^C).

The forecast combinations are considered only on the X^C forecasting because it was identified that their use allows us to obtain the best specific predictions. The methods used in the combination procedure are: Smoothing Splines, Auto Arima, the mean of the period, the Theta method and the Exponential Smoothing. In this way these different methods return forecasts, and the combined forecasts can be considered as the average of the models. As a prediction interval we use an interval of 5 periods. It is interesting to note that the combination approach does not produce better results in every case in comparison with one method (the best one). At the same time as the predictions from the forecasts in the X^C and Y^C attribute time series we can expect to capture the evolutionary features of the beanplots. It is clear that the predictions can have relevant consequences also for the associated risk analysis (an increasing difference between beanplot extremes that can indicate a higher risk).

8.6.1 Combination Schemes

Various schemes of combination in forecasting can be used:

1. Equal weighting (average) on the different methods SETAR¹⁸, KNN¹⁹, Regime Switching²⁰, etc.
2. Different weighting proportional to the quality of the forecasts²¹
3. Different weighting considering a regression approach (a Granger Ramanathan Approach: see Granger Ramanathan 1984 [322]). In this case it could be useful to consider different regressions like the Nonlinear Least Squares NLS²² or the genetic algorithms²³
4. Hybrid approaches to determine optimal decomposition of the series and use them in a combination²⁴
5. Ranking of the different forecasts²⁵.

Following Timmermann 2006 [658] and Andrawis Atiya El-Shishiny 2010 [18] let f be a forecast of a single so in all cases we have a set of n forecast f at an interval t so we can combine.

f is a forecast of a single model:

$$F_t^{CM} = w_1(t)f_1^1 + w_2(t)f_2^2 + w_3f_3^3 + \cdots + w_mf_t^m \quad (8.9)$$

With: $w_1 + w_2 + w_3 + \dots + w_n = 1$ and $w(1) \geq 0, w(2) \geq 0 \dots w(n) \geq 0$ for all the weights w .

1. Base scenario: $w_1 = w_2 = \dots w_n = 0$ considering all forecasts

¹⁸Tong 1990 [665] Tong 1983 [664] Tong Lim 1980 [663]

¹⁹Yakowitz 1987 [709]

²⁰Hamilton 2005 [334]

²¹Timmermann 2006 [658] Andrawis Atiya El-Shishiny 2010 [18]

²²Fox 2002 [272]

²³Rousseeuw and van Driessen (2006) [598]

²⁴Zhang 2003 [717] and Maia, De Carvalho and Lurdermir 2006 [477]

²⁵Kisimbay 2010 [425]

8.6. The Forecasts Combination Approach

2. Alternative: $w_1...w_n = 0$ discarding some forecasts $f_1...f_n$, using the trimmed mean
3. Alternative: $f_{t+h} = \sqrt[n]{f_{t+h}^1 + f_{t+h}^2 + ... f_{t+h}^n}$, using the geometric mean. See Andrawis Atiya Shishiny (2010) [18] and Timmermann (2006) [658].
4. Alternative: using the harmonic mean: $f_{t+h} = \frac{n f_{t+h}^1 f_{t+h}^2 ... f_{t+h}^n}{f_{t+h}^1 + f_{t+h}^2 + f_{t+h}^n}$
5. Weight determination: $w_1...w_n \geq 0$, with: $w_1 + w_2 + w_3 + ... + w_n = 1$

8.6.2 Optimal weight determination

At the same time the weights of the combination can be differently defined²⁶:

1. $w_1 = w_2 = ...w_n = 0$ as simple criteria has the advantage of simplicity, but clearly different alternatives could be chosen. A second approach can be related to the variance (we report here the case of two weights as an example):
2. $f_{t+h} = [f_{t+h}^1]^{w_1} + [f_{t+h}^1]^{w_2} + ... + [f_{t+h}^n]^{1-w_1+w_2}$ optimal schemes using a search algorithm using the geometric mean
3. Let two variances σ_2^1 and σ_2^2 represent the covariance between the two forecasts. In this case, assuming the weight sum is 1 each weight can be obtained by: $w_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_{21} + \sigma_{22} - 2\sigma_{12}}$ and also $w_2 = 1 - w_1$

²⁶Timmermann 2006 [658] Andrawis Atiya El-Shishiny 2010 [18] Kang (1986) [417] Deutsch Granger Terasvirta 1994 [196]

4. or also weighting by MSE (Stock Watson 2004 [644]), so we

$$\text{obtain: } w_1^h = \frac{\sum_{j=-k}^k MSE_{h+j}^2}{\sum_{j=-k}^k MSE_{h+j}^1 + \sum_{j=-k}^k MSE_{h+j}^2} \text{ and also}$$

$$w_2^h = \frac{\sum_{j=-k}^k MSE_{h+j}^1}{\sum_{j=-k}^k MSE_{h+j}^2 + \sum_{j=-k}^k MSE_{h+j}^1}$$

8.6.3 Weight determination by regression

Another possibility is to directly estimate the weights by regression, having the different forecasts as independent variables and the true values as dependent.

In this case it is also possible to define various options at the same time:

1. Linear weight parameterization: Granger and Ramanathan Combination²⁷ estimate omitting the intercept: $y_t = \beta_0 f_1 + \beta_1 f_2 + \dots + \beta_n f_n + \epsilon_t$
2. Nonlinear weights parameterization
3. Changing weights over the time²⁸

8.6.4 Identification of the components to model

At the same time it is important to identify the eventual components of the models.

1. Assumption that the series is composed of parts: for example $y_t = L_t + N_t$ where L_t is a linear autocorrelation structure and N_t is a non-linear component.

²⁷Granger Ramanathan (1984) [322]

²⁸Deutsch, Granger, Terasvirta 1994 [196]

2. Number of components to decompose the series²⁹
3. Identification of the structured parts to decompose the series³⁰.

8.6.5 Identification and implementation of the Hybrid modelling strategy

At the same time it is possible to differently identify and implement the models by considering a hybrid modelling strategy:

1. Methods choice, for example ARIMA, as in the case of Maia, De Carvalho, Ludermir (2006) [477] to estimate L_t then obtain $\epsilon_t = y_t - \hat{L}_t$ and use other methods, for example neural networks to capture the nonlinear structures of the y_t . In particular $\epsilon_t = \phi(\epsilon_{t-1}, \epsilon_{t-2} \dots \epsilon_{t-3}) + k_t$. The single forecast f_n from the model is: $\hat{y}_t = \hat{L}_t + \hat{N}_t$
2. Diagnostics
3. Combination Strategy. Combine f_1, f_2 etc.

Combinations using different assumptions

1. Different Hybrid Methods f_1, f_2 can be used, based on different assumptions or hypotheses (interval data)
2. Combination strategy. Combine f_1, f_2 and so on.

²⁹Hendry Hubrich 2006 [355] and Hendry Hubrich (2010) [356]

³⁰Hendry Hubrich 2006 [355] and Hendry Hubrich (2010) [356]

8.6.6 Using Neural Networks and Genetic Algorithm in the modelling process

Neural Networks³¹, were used in literature to improve the forecasting process of interval data (in particular see Muñoz, Maté, Arroyo, Sarabia (2007) [522] García Ascanio and Maté 2010 [291] and Maté and García Ascanio 2010 [494]). At the same time there are many studies that use neural networks in stock forecasting (see for example Lawrence 1997 [451] and Weckman et al 2008 [695])

So an option can be to use this approach in the forecasting of the attribute time series. In this work, in particular as a learning scheme, the Multilayer Perceptron was used to forecast some attribute time series (mainly the Y^C) that present a sinusoidal trajectory. The procedure was useful in analysing attribute time series which show these characteristics.

The genetic algorithm can be used in two different contexts in order to optimize predictions both as single forecast, or in forecast combination schemes.

First of all we can use the genetic algorithm to explore the best models we can define where it could be particularly difficult to model the data (for example in the case of the Y^C coordinates of the beanplot data). In that sense the Genetic Algorithm can help to identify the model using a Symbolic Regression (for the entire procedure see Schmidt Lipson in various works, for example: [608] , [610], [612] and [614]). However it should be stressed that this process is very useful in conditions of attribute time series characterized by relevant noise.

Secondly we can use the genetic algorithms to optimize the models in the case of forecasts combinations. We can consider the genetic algorithms in the regression to estimate the parameters of the combi-

³¹The literature in this area is enormous. An introduction can be found in Witten Frank 1999 [704] and Fabbri Orsini (1993) [258]

nation model, avoiding the effects due to outliers or a strange value.

In this respect the results obtained can be more robust, whilst the robustification of the procedure can improve the final forecasts (see Wildi 2007 [700]).

8.7 The Search Algorithm

There is a growing literature on finding the adequate estimation window: see Pesaran Timmermann 2007 [559] and Pesaran Pick 2010 [557]. The search algorithm is useful in detecting the best information set for the prediction. So the forecast processes are repeated to consider all the possible sets of information, where the model minimizing some statistical criteria is selected (for example we use the MAPE see: Hyndman 2006 [379]). In this way we obtain the best forecasting interval for the considered model. So the computational tool to improve the forecasts is considered to be the algorithm for the selection of the best observations (or the best information set) in the forecasting models. See Hendry (2006) [354], and Castle et al (2007) [122], and Fawcett and Hendry (2007) [266], at the same time a methodology that could be applied to forecasting in real time is in Pesaran Timmermann (2004) [558].

8.8 Crossvalidating Forecasting Models

It is important to note that the initial data are divided into two distinct sub-periods. A first one is considered as a training set, in which the different models are estimated for each attribute time series (both X^C and Y^C). A second sub-period is used to compare the results ob-

tained for different models (validation set). Various schemes³² can also be considered by taking into account relevant different periods of the series (the financial crisis for example, or other relevant situations).

This approach has the specific aim of improving the forecasts in the forecasting process, and it could be considered as a cross-validation approach for the model selection (for a review in this sense see Arlot Celisse 2010 [26]).

At the same time when a specific model is chosen it could be important to optimize the prediction by considering the optimal set of information to use.

In this case we cross-validate the different models chosen in the first phase by using the search algorithm.

8.9 Extremes and Risk Forecasting

A useful application of the beanplot forecasting over time could be to forecast the VaR, or the Value at Risk³³, as an important measure of Financial Risk, by considering both the lowest coefficient in the mixture or the lowest coordinate as descriptor.

This operation of using the prediction of Histogram Data (as a symbolic data or as aggregate representation for predicting the VaR) is proposed by Arroyo et al. 2011 [34] and Arroyo, González-Rivera, Maté, Muñoz San Roque [39] in 2010.

In that sense, the VaR forecasting corresponds to predicting the lowest values of the beanplot data over time in the specific temporal interval. As expected, the temporal interval needs to be the same for the VaR computations.

Various approaches can be considered as well in literature.

³²Friedman Hastie, Tibshirani 2009 [282] Giudici 2006 [312] and Refaeilzadeh, Tang and Liu 2009 [570]

³³Jorion 2006 [414], Holton 2003 [363]

An approach is related to the VaR modelling in Engle and Manganelli 2004 [252] using the quantile regression (see in this respect Koenker and Basset 1978 [430]). An interesting comparative analysis by considering different approaches is the analysis performed by Andre Nogales and Ruiz 2009 [19]. Another possibility is using the Copulas (for an approach to Copula Theory see Nelsen 1999 [532]).

Another important task in measuring the risks, by considering the financial risk, is to predict the extremes. A first approach is that of directly using Extreme Value Theory in a context of regression or modelling, this can be found in Toulemonde et al. 2009 [666].

8.10 Beanplot Forecasting: Usefulness in Financial Applications

Beanplot forecasting can be useful in many different contexts, for example in statistical arbitrage and trading. Various different contexts can be considered in that sense. For example, pair trading can be explicitly done by considering beanplot data. In fact, by identifying a couple of similar stocks it is possible to predict the dynamics in the long run and to profit from the differences. At the same time various other strategies can be considered, for example identifying some patterns in data as seasonalities which could be exploited by considering some seasonal trading strategies. Other clear relevant applications of the beanplot forecasting are in the field of Risk Management, in fact by considering the dynamics of the stocks using the maxima information available it is possible to predict the outcome of different events in a more consistent way. Study events in this sense can be used to detect possible outcomes related to the single financial event.

Possible applications: short term forecasting: trading and statis-

tical arbitrage, long term forecasting: macroeconomic and financial analysis, quantitative Models for forecasting economic and financial variables, Tactical Asset Allocation, Risk Management, Scenario Analysis and Stress Testing.

Summary Results: Forecasting
In forecasting we consider the forecasts of the TSFA model in the model-based coefficient estimation. In the second type of approach we forecast the attribute time series.
Various different approaches can be considered in the forecasting process, all the approaches need to be based on an identification of the external model to adopt.
A combination of external models could be very useful if it is possible to find a group of forecasting models which perform well. In this case, with the combination we reduce the uncertainty of choosing a unique model and we consider eventual parameter drift.
In the forecasting procedures, we can use the search algorithm to improve the forecasts by choosing the optimal set of information to include in the model.

Chapter 9

Beanplots Time Series Clustering

A clustering problem is finding similar subgroups of specific items in a specific set by assuming absence of other information (Jain 2010 [402] and Jain Murty Flynn 1999 [403] Xu Wunsch and others 2005 [708]). In particular the single observations which are the most different between the subgroups and the most similar between each other need to be considered. There are relevant cases in which the interest is not only in analysing one series but groups of series, for example in the case of the study of the markets or portfolios (Pattarini Paterlini Minerva 2004 [551]). In practice, in these cases, some time series tend to respond to asymmetric shocks similarly (Basalto and De Carlo 2006 [62]). So it is interesting to discriminate the different behaviors over time. There are other cases in which the clustering process is not straightforward but the series need to be preprocessed. This is the case of complex time series related to phenomena like the financial markets (see Sewell 2008 [619] for a review of the characteristics of these series). In this part we deal with complex time series, represented as beanplot time series (BTS).

We start from a complex financial time series to obtain the associated beanplot time series (BTS). In particular we start from a series $\{y_t\}, t = 1 \dots T$ with $y_t \in \mathbb{R}$ and we can have a single value in \mathbb{R} . In the case of the high frequency data we have a huge quantity of observations that need to be aggregated to handle them adequately. At the same time high frequency data present irregularities due to the specific nature of the data (Engle Russell 1998 [253]. In this case we are dealing only with the mean and so we face a loss of information.

Various proposals exist to take into account the internal representations which consider the entire data structure and also the data centers. There are various proposals in the literature: those dealing with intervals, histograms (in Arroyo et al. 2011 [38] and 2010 [39]) and those dealing with distributions.

All these proposals are related to the field of Symbolic Data Analysis where interval and histograms are particular cases of symbolic data (see Diday Noirhomme 2008 [218] and Billard and Diday 2003 [85]). In Clustering it is very important to retain the relevant data information, for this reason we consider the beanplot because it allows the retention of the information on the data structures (Drago Lauro and Scepi [235] and Drago Scepi 2010 [237]. The density trace in particular gives the chance to take into account structural changes that can occur. In practice, Beanplot time series (BTS) can represent the features of the initial time series (the beanline corresponds to the aggregated time series). In particular, beanplot can represent at the same time trend, cycle and even structural changes of the original time series. Various clustering approaches for beanplots can be defined:

1. Model approach: Single beanplot or entire time series (Romano Lauro Giordano distance¹) see Chapter 9.1
2. Time Series of Attributes (minima, maxima) (Correlation, Cep-

¹Romano Giordano Lauro 2006 [594]

9.1. Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

- stral, Time Warping Distance) Chapter 9.2-9.3
- 3. Time Series factors of trajectories (Correlation) see Chapter 9.2-9.3
- 4. Model Based Clustering (Model Based) see Chapter 9.4-9.5
- 5. Clustering Beanplots Data with Contiguity Constraints Chapter 9.6
- 6. Single Beanplot (Wasserstein and Euclidean distance) see Chapter Chapter 9.7
- 7. Building Beanplot Prototypes (BPP) using Clustering Beanplot Time Series (BTS) Chapter 9.9
- 8. Ensemble Clustering see Chapter 9.10

The first method is analyzed in Chapter 9.1 and the others in Chapter 9.2-9.10. The different methods reflect the differences in the approaches seen in the internal modelling of the beanplot time series (BTS).

9.1 Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

Internal Models in this case are sequences of mixture coefficients. In order to cluster multiple time series of beanplot B_{y_t} with $t = 1 \dots T$ we start from a synthesis of such multiple time series obtained by a multiple factor time series approach $\xi_{v,t}$. In particular we use a suitable

distance between models, combining a convex function of the differences in model coefficients with corresponding fitting indexes (Romano Giordano Lauro 2006 [594]).

In order to cluster multiple time series of beanplot B_{y_t} with $t = 1 \dots T$ starting from the model synthesis, the final aim is to recognize similar factor time series with similar dynamics. Thus we obtain the factorial time series $\xi_{v,t}$ and we compute the dissimilarity matrix. So we use an adequate distance, as for example, the distance from models in Romano, Giordano Lauro (2006) [594] where we are trying to consider the dynamics of the different factors.

In order to cluster a set of beanplot time series (BTS) B_{y_t} with $t = 1 \dots T$, we use a suitable distance between models that combine a convex function of the differences in model coefficients with corresponding fitting indexes (Romano Giordano Lauro 2006 [594]). Following Signoriello 2008 [630] the two pieces of information are combined to define the following measure:

$$IM(p^j, p^{j'} | \lambda) = \lambda IM_P + (1 - \lambda) IM_R \quad (9.1)$$

with $\lambda \in [0, 1]$. The IM measure is a convex combination of two quantities IM_P and IM_R , where IM_P is the L_2 -norm between the estimated coefficients:

$$IM_P = \left[\sum_{k=1}^{K-1} \left(w_k^j - w_k^{j'} \right)^2 \right]^{\frac{1}{2}} \quad (j \neq j') \quad (9.2)$$

and IM_R is the L_1 -norm between the chi square:

$$IM_R = \left| w_K^j - w_K^{j'} \right| \quad (j \neq j'). \quad (9.3)$$

In the clustering process of beanplot time series (BTS) (

The two pieces of information are combined to define the following measure:

9.1. Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

$$IM_T(p_t^j, p_t^{j'} \dots p_T^j, p_T^{j'} | \lambda_t \dots \lambda_T) = \sum_{t=1}^T \lambda IM_{P_t} + (1 - \lambda_t) IM_{R_t} \quad (9.4)$$

with $\lambda \in [0, 1]$. The IM_T measure is a convex combination of two quantities IM_{P_t} and IM_{R_t} , where IM_{P_t} is the L_2 -norm between the estimated coefficients:

$$IM_{P_t} = \left[\sum_{k_t=1}^{K_t-1} \left(w_{k_t}^j - w_{k_t}^{j'} \right)^2 \right]^{\frac{1}{2}} \quad (j \neq j') \quad (9.5)$$

and IM_{R_t} is the L_1 -norm between the chi square:

$$IM_{R_t} = \left| w_{K_t}^j - w_{K_t}^{j'} \right| \quad (j \neq j'). \quad (9.6)$$

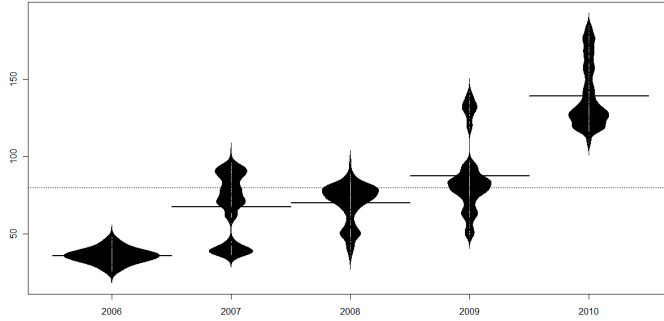


Figure 9.1: Amazon (AMZN)

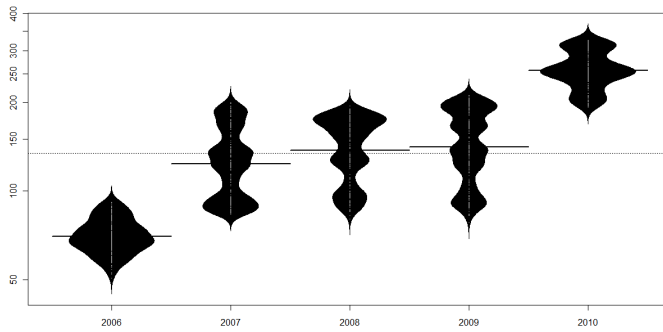


Figure 9.2: Apple (AAPL)

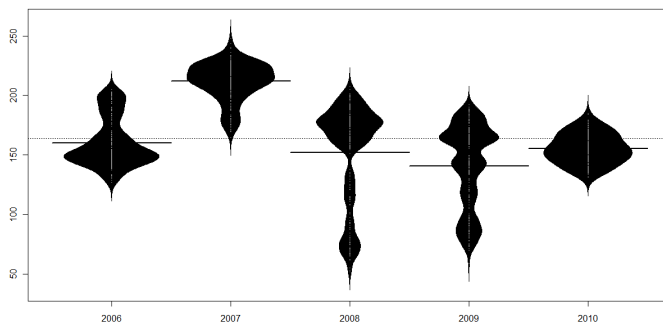


Figure 9.3: Goldman Sachs (GS)

9.1. Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

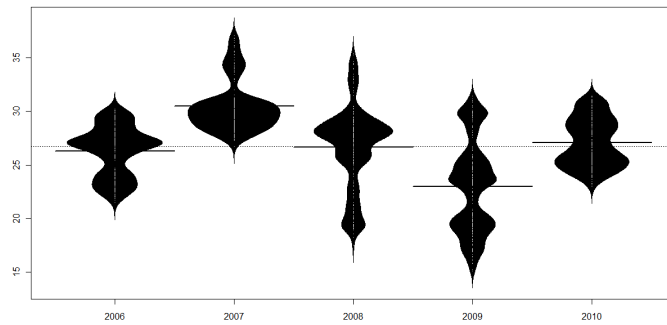


Figure 9.4: Microsoft (MSFT)

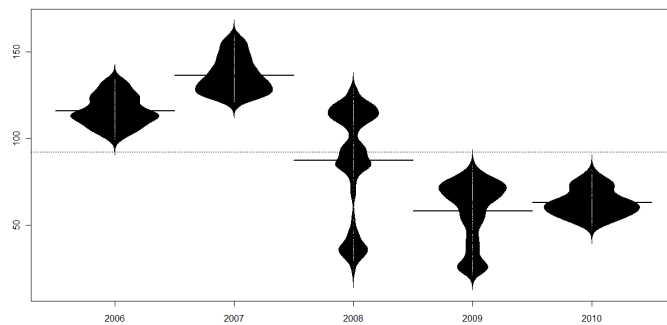


Figure 9.5: Deutsche Bank (DB)

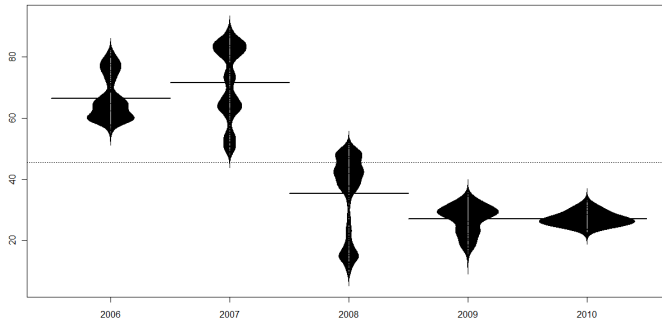


Figure 9.6: Morgan Stanley (MS)

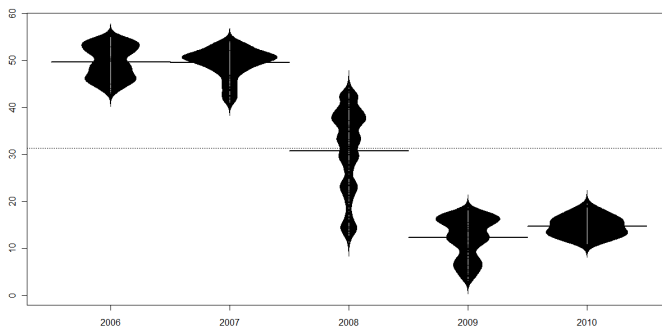


Figure 9.7: Bank of America (BAC)

9.1. Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

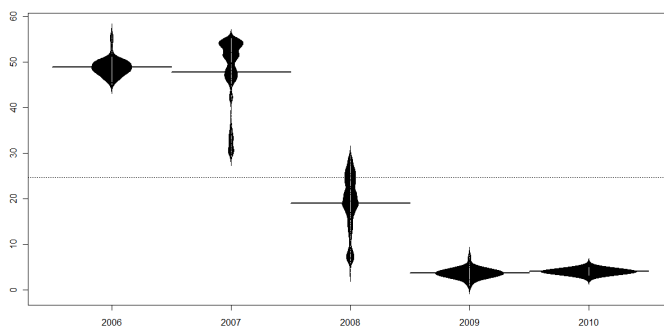


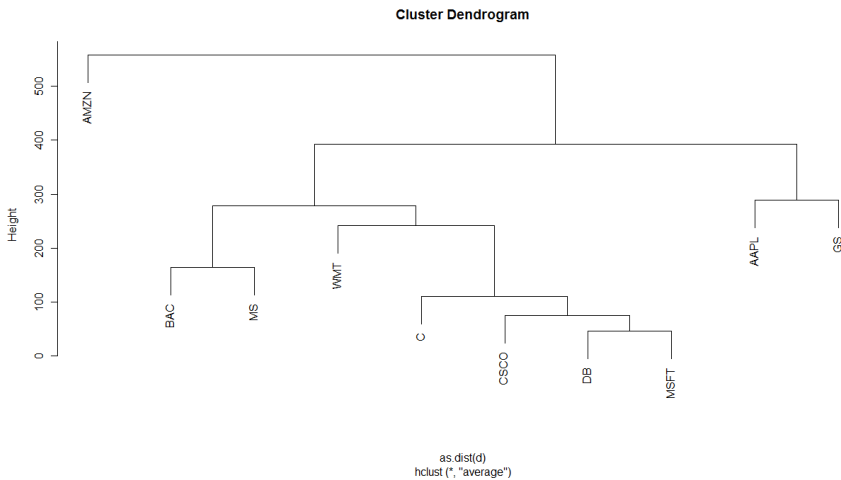
Figure 9.8: Citigroup (C)

9.1.1 An application on real data: Clustering stocks in the US Market

Now we begin to consider a portfolio of stocks, in which we study its behavior as Beanplot Time Series (BTS) (figure 9.1, figure 9.2, figure 9.3, figure 9.4, figure 9.5, figure 9.6, figure 9.7, figure 9.8). All the stocks are related to the DJI Market. We firstly visualize the series, then we represent the stocks by considering them in four different subperiods. We estimate the coefficient p and we use the model distance to obtain the dendrogram. We consider comparatively four different subperiods and the entire period, using the appropriate distance. It is interesting to note that the method discriminates the different profiles of the different stocks. In practice we consider and compare the risk profiles for four periods. In doing so we see that one company performs as an outlier (Amazon) because they are not so affected by the crisis (which is first diffused by means of financial linkages). In that sense we can observe that financial companies tend to present a very

similar risk profile and to be clustered in the same group (in particular Morgan Stanley MS and Bank of America BAC). The dendrograms are presented in figures figure 9.9, figure 9.10, figure 9.11, figure 9.12, figure 9.13 .

Figure 9.9: Dow Jones Market



Clustering the Beanplot Time Series (BTS): DJI Market subperiod 2007-2008

Clustering the Beanplot Time Series (BTS): DJI Market subperiod 2008-2009

Clustering the Beanplot Time Series (BTS): DJI Market subperiod 2009-2010

Clustering the Beanplot Time Series (BTS): DJI Market subperiod 2010-2011

9.1. Clustering Multiple Beanplot Time Series (BTS): the Model Distance Approach

Figure 9.10: Dow Jones Market 2007–2008

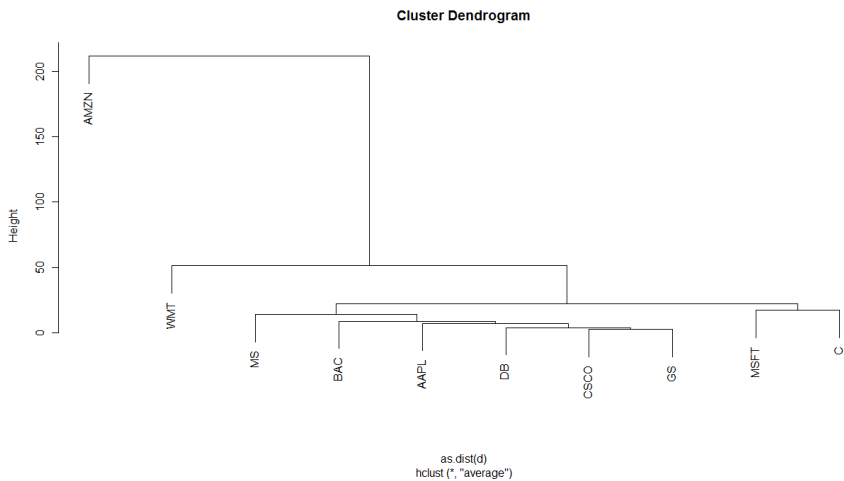


Figure 9.11: Dow Jones Market 2008–2009

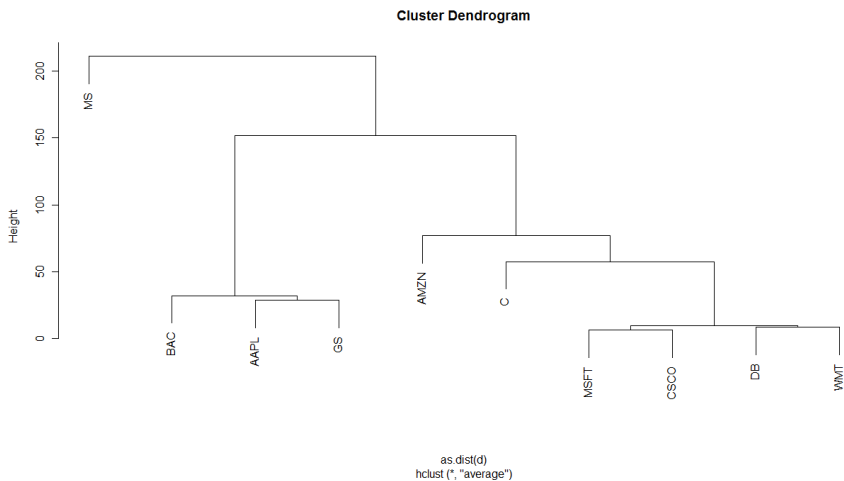


Figure 9.12: Dow Jones Market 2009–2010

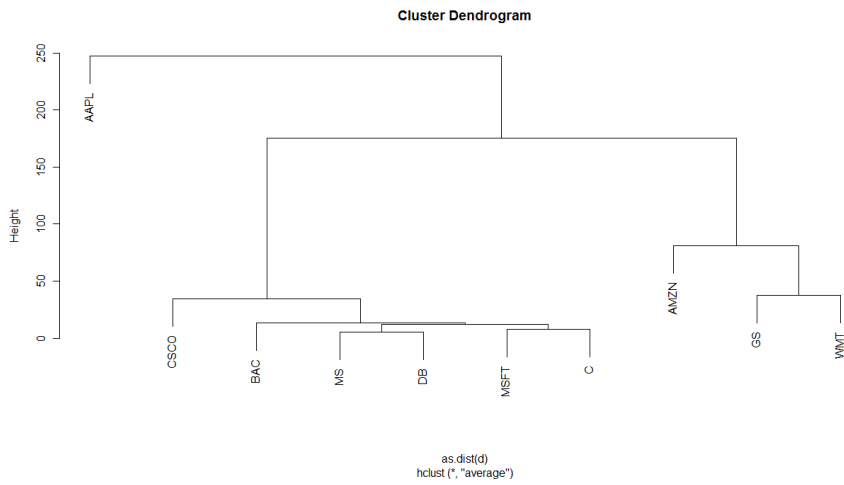
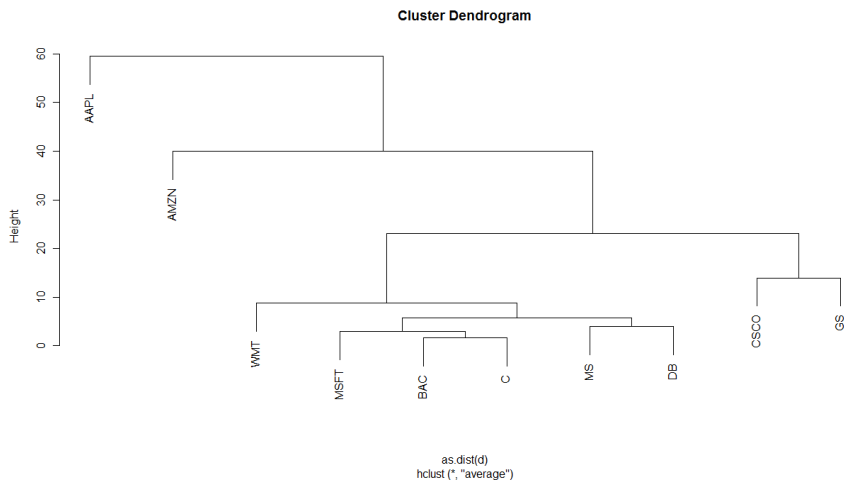


Figure 9.13: Dow Jones Market 2010–2011



9.2 Internal Modelling and Clustering: the Attribute Time Series Approach

In the internal modelling phase it is relevant to choose the bandwidth h for the beanplot time series (BTS) and the number of descriptor points or features considered n . We need to select the h parameter for the entire beanplot series (in order to visualize the series there is no need to choose the h parameter selected by the Sheather-Jones method). The choice is related to the structure of the data in each temporal interval t . The feature n represents both the data structure of the interval temporal and the beanplot evolutive dynamics over time by its attribute time series. In particular, we have to characterize the beanplot by considering either the X^C and the Y^C as coordinates. In that way we obtain the internal model of the single beanplot at the interval t . As output of the internal modelling process we obtain the attributes time series for the X^C and the Y^C of the beanplot time series (BTS). By deciding the number of features n to take into account (for example for X^C the 25, 50 and 75th quantile and the related Y^C coordinates (Drago Lauro Scepi 2009 [233]) we have a higher number of attribute time series). We choose these coordinates without considering the extreme values that could be affected by outliers. The crucial point in the internal modelling process is the choice of the bandwidth h and the choice of the number of the n features, considering them in relation to the data structure. It is important to validate the internal model (the n choice of the features and the chosen bandwidth h) and its adequacy to represent initial data. In general it is necessary to take into account a lower bandwidth h in the series with a higher number of features n if there is a higher level of observations: thus, we can capture a higher number of features. In particular, we consider that the higher the complexity of the original time series the higher the complexity requested and the number of features to be taken into

account. Therefore, it is necessary for most of the cases to reproduce the structure of the beanplot by considering at least three descriptors for the X^C and three descriptors for the Y^C .

9.3 Classical Approaches in Clustering Beanplot Features

A first approach, if we are interested in one or more specific attribute time series (for example the attribute of the minima to compare the behavior in crisis), is that of directly clustering the attribute series of the different beanplot time series (BTS) (Algorithm 12). In this sense classical distances can be used, like the Correlation distance, the Cepstral distance, and the Time Warp: see Liao 2005 [455].

At the same time if we need to cluster the entire beanplot time series (BTS) we need to synthesize and to model (the external model) all the attribute time series. In particular we use the time series factor analysis (TSFA) depicted in Meijer Gilbert 2005 [500] and Gilbert Meijer 2006 [306] to estimate the attribute time series both for the X^C and for Y^C . We start from the n observed processes $a_{i,t}$ with $i = 1..n$ and $t = 1..T$, where we search from the k unobserved processes (the factors) $\xi_{i,t}$ with $t = 1..T$ and $i = 1..k$ and we obtain the measurement model:

$$a_t = \alpha + V\xi_t + \epsilon_t \tag{9.7}$$

Where α is a vector of intercepts, V is an n, k matrix of factor loadings and ϵ is an n vector of random errors. Each factor score represents a measurement model of a latent variable that is the underlying phenomena of interest. At the end of the procedure we obtain a set of Factors for each attribute time series. For the measurement of the factor score predictor we use the Bartlett predictor, following Meijer

9.3. Classical Approaches in Clustering Beanplot Features

Gilbert 2005 [500], Wansbeek and Meijer. (2000) [690] and Wansbeek Meijer (2005) [689].

$$\xi_t = (V^t \Psi_t^{-1} V)^{-1} V^t \Psi^{-1} V^t (z_t - \alpha_t) \quad (9.8)$$

Where: $\psi_t = Cov(\epsilon_t)$. The loadings can be estimated by FA estimators (ML in particular) using the sample covariance of the error (Meijer Gilbert 2005 [500]). We compute one factor time series either for the X^C or for the Y^C using the attribute time series. Either one of the two factorial time series represents the X^C and Y^C , where the first one represents the general dynamics of the beanplot (in particular the location and the size) and the second one represents the response of the shock or the short run dynamics (the shape).

The final aim of the clustering process (Algorithm 13) is to recognize groups of time series with a synchronous dynamic (related to the location of the beanplot or the X^C) and a similar response to the shocks (related to the Y^C). Having obtained for each beanplot the factorial time series both for the X^C and the Y^C , we compute the dissimilarity matrix. In particular we use three versions of the distance known in literature as correlational distances (see for example Dose and Cincotti 2005 [225]), where we try to specifically recognize the correlation between the dynamics of the different synthesizing factors over time. So we have:

$$d(Y, X) = 1 - (c_{\xi_{bY_t}, \xi_{bX_t}}) \quad (9.9)$$

Where: $Y = (\xi_{bY_1}, \xi_{bY_2}, \dots, \xi_{bY_t})$ and $X = (\xi_{bX_1}, \xi_{bX_2}, \dots, \xi_{bX_t})$, and $c_{\xi_{bY_t}, \xi_{bX_t}}$ is the correlation coefficient between the two factorial time series related to X^C and Y^C . We use as well² other correlational distances to compare the results:

²See Glynn (2005) [751]

$$d(Y, X) = \frac{1 - (c_{\xi_{bY_t}, \xi_{bX_t}}))}{2} \quad (9.10)$$

Considering the absolute values:

$$d(Y, X) = 1 - |(c_{\xi_{bY_t}, \xi_{bX_t}})| \quad (9.11)$$

and finally:

$$d(Y, X) = \sqrt[2]{1 - (c_{\xi_{bY_t}, \xi_{bX_t}})^2} \quad (9.12)$$

We obtain the dissimilarity matrices related and we use different clustering methods to compare the different results. We use the hierarchical clustering and the non hierarchical clustering by applying different methods. At this point we can apply different methods to observe the robustness of the results.

9.3.1 Application: classifying the synchronous dynamics of the world indices beanplot time series (BTS)

The methods depicted in the previous chapters are transformed in R programs and experimented both on simulated data to test the characteristics of the methods, and also on real data. In the application on real data we consider various time series related to the index of the major stockmarkets around the world. In particular we consider 14 markets related to different continents. The interval period considered, for the collected data, is the period 2001-2008. As a first step we compute the beanplot time series (BTS) from the original time series, and we use the information of the bandwidths to explore the volatility (represented over time). Secondly, we identify the bandwidth and we compute the attribute time series for all the beanplots both for the

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$
Result: A vector with n elements assigning each time series of attributes for X^C and Y^C descriptors to each k group

```

begin
    Choice of the  $I$  interval temporal to use
    Choice of the  $n$  points to represent
    Choice of the  $h$  bandwidth to use
    for  $t \in T$  do
        Estimating the  $X^C$ 
        Estimating the  $Y^C$ 
        Clustering the attributes time series for  $X^C$  and  $Y^C$  for
        each beanplot time series (BTS)  $\{b_{Y_t}\} t = 1...T$ 
    end
    Are the clusters fitting the data adequately?
    if the clustering method is not adequately fitting then
        change the number of descriptor points  $n$  or the
        bandwidth  $h$ 
    end
end

```

Algorithm 12: Beanplot clustering: attributes

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$

Result: A vector with n elements assigning each factor time series of attributes for X^C and Y^C descriptors to each k group

begin

Choice of the I interval temporal to use

Choice of the n points to represent

Choice of the h bandwidth to use

for $t \in T$ **do**

Estimating the X^C

Estimating the Y^C

Estimating the factor time series (BFT) for each
beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$

Clustering the factor time series for X^C and Y^C for each
beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$

end

Are the clusters fitting the data adequately?

if *the clustering method is not adequately fitting* **then**

change the number of descriptor points n or the
bandwidth h

end

end

Algorithm 13: Beanplot Clustering: factor time series (BFT)

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1 \dots T$

Result: A vector with n elements assigning each time series of attributes for $p_1 \dots n$

begin

Choice of the I temporal interval to use

Choice of the n mixtures to represent

Choice of the h bandwidth to use

for $t \in T$ **do**

 Estimating coefficients p

 Clustering the Beanplot time series (BTS) using the model distance

end

Is the model fit adequate?

if *the model fit is not adequate* **then**

 change the temporal interval I , number of mixtures n or the bandwidth h

end

Are the clusters fitting the data adequately?

if *the clustering method is not adequately fitting* **then**

 change the number of parameters n or the bandwidth h

end

end

Algorithm 14: Beanplot clustering: mixtures

X^C and the Y^C . We consider three specific time series both for the X^C and for the Y^C . At this point we consider both the true different groups of time series and we synthesize the time series regarding only one factor for the X^C and one factor for the Y^C . We obtain the factor time series. It is interesting to note the strong impact of the financial crisis on all the world stockmarkets jointly, where there are various different responses due to the different economic policies. At this point we classify the beanplot time series (BTS) and analyse the synchronous behavior to compute the dissimilarity matrix, by using different distances. Finally, using different methods we compare the results obtained from the hierarchical and non hierarchical clustering process. We retain a factor 1, that could be interpreted according to size and data location, which determines classification due to the different development characteristics of the different markets. Where Brazil and Mexico represent the "developing markets", other markets can be considered "developed" such as the US market. Moreover the developing markets show a better response to the crisis (and they finish in the same cluster with these characteristics) as we can observe from the factorial time series. More in particular, for the developing markets, there is higher growth and instability in the long term but a better response to the financial crisis in the short term. It is interesting to note a quick reaction to the crisis by the US and Japan due to the expansionary economic policy and general policies adopted. Hong Kong reacts as a specific market (in particular through its ties to China) and Sweden represents an isolated market. The second factor Y^C represents the dynamics of the beanplot shape, and we observe from the different clusters that there are similarities in the shock responses (representing also the beanplot shape). Markets that are near through their close proximity, in that sense, show similarities in behaviour due to the financial connection. We expect that markets close in proximity tend to behave similarly, and infact we observe that the clusters present regroupments related to the European markets, the

Asian markets and "signal markets" such as Japan and the US. In particular the data seems to show that in the short run the geographical factor seems to be fundamental for defining groups for Y^C . The short run dynamics is strongly related to the shocks and the contagion mechanisms. The geographical spread of the shocks occurs if we consider financial ties between different markets. For France and Switzerland, in particular, we observe a strong reaction to the shocks and synchronous movement of the markets whereas for the European countries we observe various levels of synchronous movements of the markets due to their common economic policies and economic shocks. Finally, for the Asian and Latin American countries we find a general instability (Brazil Mexico), and a connection between shocks in Asian countries (Indonesia, Hong Kong, Singapore etc.). The instability of the Latin American countries tends to have an impact also on the Spanish market. Finally, Japan and the US are markets that represent "signals" (which means they tend to behave differently compared to other markets).

9.4 Model Based Clustering and Modern Framework

In the first part of the work we have considered classical clustering methods based on specific heuristics. Finite Mixture model structures can be used as well in the clustering process, where, in particular, each single component distribution can be considered a cluster. For a review of these methods in the context of clustering see in particular Melnykov Maitra (2010) [502]. Finite Mixture Models provide the basic foundation for a different approach: the Model Based Clustering (Wolfe 1963 [705]). In this new framework, Bock 1996 [94], Fraley Raftery (2002) [277] and Bock 1998 [92] has proposed a clustering ap-

proach based on probability models. Clusters, in this approach, can be considered a component probability distribution (Fraley Raftery 2002 [277]) where the foundation is a formal statistical approach (Fraley Raftery 2007 [279]). Heuristic approaches can be used as approximate estimation methods for some probability models (Fraley Raftery 2002 [277]). This second clustering strategy is necessary because of the data complexity of the behaviours of the original data. In the first approach, that is considering the TSFA analysis, we synthesized the information, here we want to obtain clusters considering the different behaviors of the different X^C and Y^C (at a lower or a higher level). In this sense we are taking into account the data complexity and we consider the different features of the beanplot behavior over time in the clustering process. In particular we consider n time series of beanplots characterized by different features $X^C - Y^C$. Each feature $X^C - Y^C$ can be seen as a single representation and could be differently modelled considering the different time series. Another important idea underlying this approach is that by considering subperiods over time we observe the change over time of the data structure and the models. So we are considering data-analysis not as a photograph of the situation but as the specific dynamics of change over time (following an approach in a different context like Riani 2004 [582]).

9.5 Feature Model Based Clustering for Beanplot Time Series (BTS)

The procedure can be defined in some sequential steps: in a common way with respect to the TSFA methodology of beanplot clustering, we had to define the set of the attribute time series. We obtain, for example (the higher the level of definition we choose the higher the number of descriptor points) six descriptor points of three X^C and

9.5. Feature Model Based Clustering for Beanplot Time Series (BTS)

three Y^C for each time. Having obtained the attribute time series we consider each couple of $X^C - Y^C$ attributes. Optionally we can choose an aggregation subperiod in which we can obtain a specific value $X^C - Y^C$ for the beanplot time series (BTS) for the entire period (or for a subperiod). We perform the model based clustering procedure for each couple of characteristics considered (Algorithm 15) in the subperiod and we obtain the clustering model and the classification for each feature. It is important to note that the expectation is that the beanplot time series (BTS) will be different in their features and characteristics over time.

In fact, some attribute series tend to perform differently over time t (for example, different minima and maxima) and beanplot features can capture these differences. As already stated in the first part of the work, in the Y^C features we are considering the variability or the short run movements, in the X^C attribute time series we are considering the mean effect over time, or the long run dynamics. Here, following Fraley Raftery 2002 [277], we perform a model based cluster analysis for the different beanplot $X^C - Y^C$ features, each considered as a temporal observation. Following Fraley and Raftery 2007 [279], we are making the important assumption that in a specific set of data, the observation z (related to the a feature $X^C - Y^C$) is generated by a mixture density:

$$f(z) = \sum_{p=1}^G \tau_p f_p(z) \quad (9.13)$$

where τ_p with $(\tau_p \in (0, 1)$ and $\sum_p^G \tau_p = 1)$ and f_p as a probability density function of the observations belong in the group p . The mean μ_p and the covariance matrix Σ_p of the component distributions have the probability density function:

$$\phi(z_i; \mu_p, \Sigma_p) = \frac{\exp \left\{ -\frac{1}{2} (z_i - \mu_p)^T \Sigma_p^{-1} (z_i - \mu_p) \right\}}{\sqrt{\det(2\pi \Sigma_p)}} \quad (9.14)$$

The likelihood for the considered data is:

$$\prod_{i=1}^n \sum_{p=1}^G \tau_p \phi(x_i; \mu_p; \Sigma_p) \quad (9.15)$$

Here, the parameters $(\tau_p, \mu_p, \Sigma_p)$ are estimated by using the EM algorithm. The mean, μ_p and the covariance matrix Σ_p characterize the different components. In particular, Covariances Σ_p fix the geometric features of the clusters, especially the shape, the volume and the orientation. In this sense, the authors, Celeux Govaert (1995) [125], Fraley Raftery (1998) [275] and Fraley Raftery (2002) [277] proposed to specifically parametrize the group covariance matrices through eigenvalue decomposition of the Gaussian Mixture Model:

$$\Sigma_p = \lambda_p D_p A_p D_p^T \quad (9.16)$$

Where in particular: D_p is the orthogonal matrix consisting of the eigenvectors, A_p is a diagonal matrix with entries proportional to the eigenvalues of Σ_p , and λ_p is an associated scalar constant to the ellipsoid volume (Fraley 1996 [274]). The geometric characteristics of the components are discovered when the parameters are found (Fraley Raftery 2007 [279] and Raftery Fraley 2007 [571]) in particular D_p the orientation of the specific component p of the mixture, A_p represents the shape, whilst λ_p represents the volume. It is important to note that we obtain different models over time. So we repeat the model-based clustering for each subperiod but also for each couple of features $X^C - Y^C$. The general estimation method of the parameters is the EM algorithm that shows the different shape of the mixtures as well (for an adequate categorization of the different models see Fraley Raftery 2002 [277]). In general the EM algorithm is a statistical tool in mixture estimation problems or also those involving missing data (Borman 2009 [103]). The EM algorithm can be considered as

9.5. *Feature Model Based Clustering for Beanplot Time Series (BTS)*

a procedure structured in two distinct parts. In the "E" part, the conditional expectation of the complete data log-likelihood is computed considering the data, with parameter estimates (following Fraley Raftery 2002 [277]). In the "M" part, the parameters maximizing the expected log-likelihood (from the antecedent "E" part) are computed. It is necessary in our context to repeat the analysis for the different beanplot features (or attributes) considered. The algorithm tends to converge if the increases of the likelihood in the sequential iterations increase (for each feature considered) Borman 2009 [103] and also Fraley Raftery 2002 [277]. At the end of the procedure we obtain the parametrizations of the different models (see also Fraley Raftery 1999 [276]). By obtaining the parameters it is possible to show the structure of the components (Fraley Raftery 1998 [275]), the different parameterizations of the covariance matrix are shown in (Fraley Raftery 2006 [278]). Moreover, where each cluster corresponds to a different statistical model (Fraley Raftery 2002 [277]), a typical model selection, the problem is to choose the number of partitions or clusters in the model based clustering process. For the model selection, various methods can be used (see Fraley Raftery 2002 [277]), one of the most known in literature is the BIC (see Fraley Raftery 1999 [276]). So we choose the model that maximizes the BIC index for every beanplot feature, because a penalty term is added to the number of parameters of the log-likelihood considered. As the authors state (Fraley Raftery 2002 [277]) there is a specific trade-off in selecting a simple model or a complex one. The second calls, usually, for a lower number of clusters.

9.5.1 **The choice of the temporal windows**

The choice of the temporal interval to be used in the dynamic part of the procedure is a decision linked to the objectives of the analysis. In particular, it can be interesting to compare the model based clustering process in two (or more) different subperiods, alternatively it

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1 \dots T$

Result: A vector with n elements assigning each time series of attributes for $p_1 \dots n$

begin

Choice of the I temporal interval to use

Choice of the n points to represent

Choice of the h bandwidth to use

for $t \in T$ **do**

Representing the n descriptor points obtaining X^C and Y^C

Model Based Clustering using X^C and Y^C

Is the model fit adequate?

if *the model fit is not adequate* **then**

change the temporal interval I , number of parameters n or the bandwidth h

end

Are the clusters fitting the data adequately?

if *the clustering method is not adequately fitting* **then**

change the number of descriptor points n or the bandwidth h

end

Is the data structure changing?

if *the data structure is changing* **then**

consider at time t a structural change

end

end

end

Algorithm 15: Beanplot Time Series (BTS) Model Based Clustering

9.5. *Feature Model Based Clustering for Beanplot Time Series (BTS)*

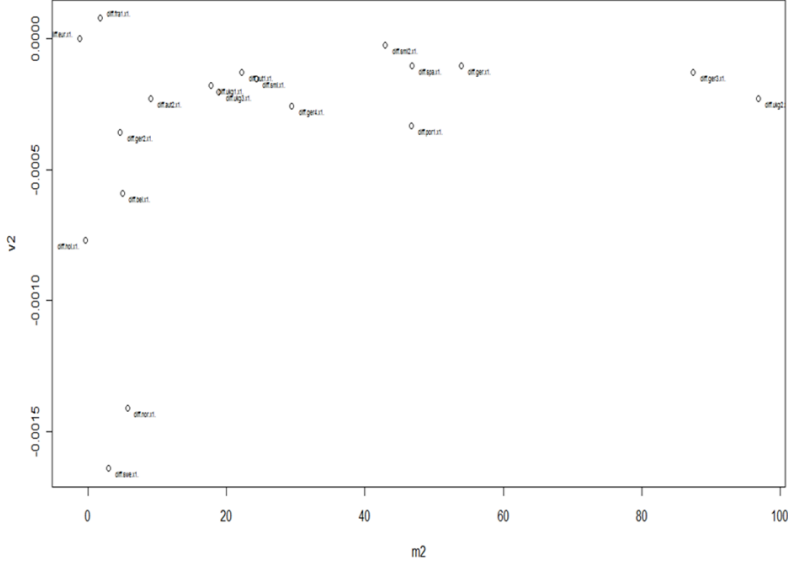
can be interesting to find a structural break in a point in the time. In particular a relevant choice is the specific interval, which determines the change of the different mixture models over time for each beanplot coefficient. In fact, each different interval can return different results in $X^C - Y^C$. It is necessary to run and compare all the different temporal intervals by considering the different results over the time. It is possible to compare, as well, the outcomes by defining different subperiods. At the same time different attribute time series can produce some outliers. So it is useful to detect them and to handle them using some specific strategies (Lipkovich Smith 2010 [463]). In this way some methods of Forward Search (Riani 2004 [582]) can be applied to the attribute time series. At the end of the dynamic analysis we obtain not only a single evaluation by a dendrogram of a "stable" situation over the time, but a specific moving image of the models and the clusters of the beanplots by considering their specific features (minima and maxima for example)³.

9.5.2 **Application: classifying the synchronous dynamics of the european indices beanplot time series (BTS)**

Symmetrically to the first part, also in this case the methods are transformed into R programs, by considering all the different phases of the process. The objective is to experiment the methods either by using simulated or real data. It is important to stress the differences between the two approaches: the classical one (Chapter 9.2-9.3) and the modern (9.4-9.5). In the classical approach we synthesize the original attribute of the beanplot and we tend to use in the analysis a higher number of series. In the modern approach we tend to select the observations by using all the features jointly to take into account the

³Atkinson Riani Cerioli 2004 [48]

Figure 9.14: Beanplot Time Series (BTS) Model Based Clustering (1)

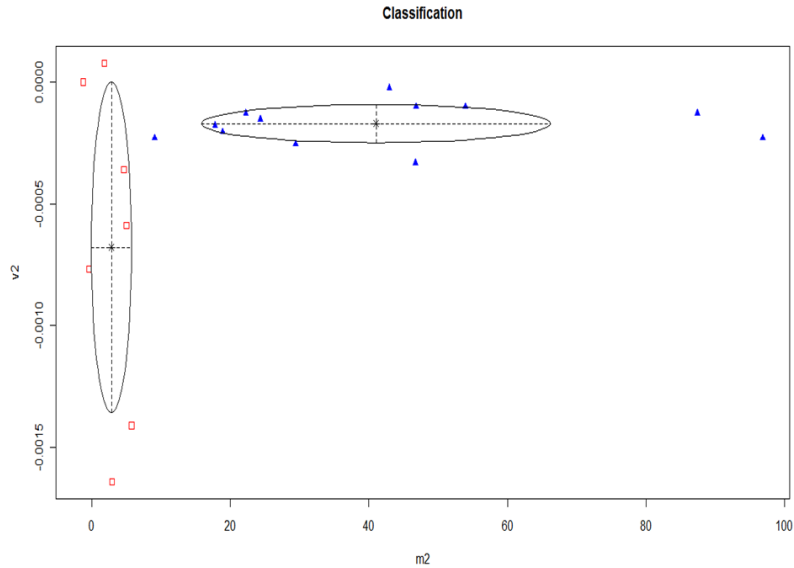


outliers.

In particular, we use here a set of data related to European Markets in which we consider the period 2003-2010. In practice we follow all the steps in the analysis described (internal and external modelling) for all the different beanplot time series (BTS). As we know, the difference we want to exploit here are related to the X^C and Y^C characteristics of the beanplots, those related to their features. Different beanplot time series (BTS) show different features in relation to the $X^C - Y^C$. We want to exploit directly these differences. In particular we consider three couples of descriptor points related to higher values, central values and lower values. It is important to note that there are relevant differences in the attribute time series, related to the complex functioning of the financial markets, single models and single charac-

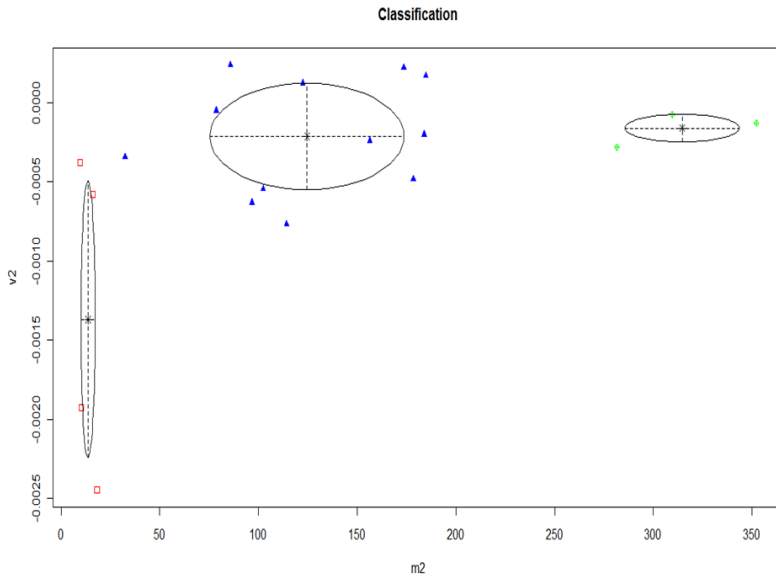
9.5. Feature Model Based Clustering for Beanplot Time Series (BTS)

Figure 9.15: Beanplot Time Series (BTS) Model Based Clustering (2)



teristics of market similarities, so we obtain different models for each feature of the beanplots. These different beanplot features can represent the complexity of the models we are considering. In particular we expect some similarities in the behavior for each period (the behavior of the influential financial areas) but at the same time some relevant differences due to the complex behavior of the series each time. This characteristic emerges in the course of the financial crisis and cannot be observed with the first method. The results are visualized in the different pictures computed running the algorithms sequentially over the time. In particular we obtain a first model based clustering analysis by taking into account the entire period (39 observations). This first analysis represents the general equilibrium of the period. It is related to the entire period, computing sequentially the results for either the

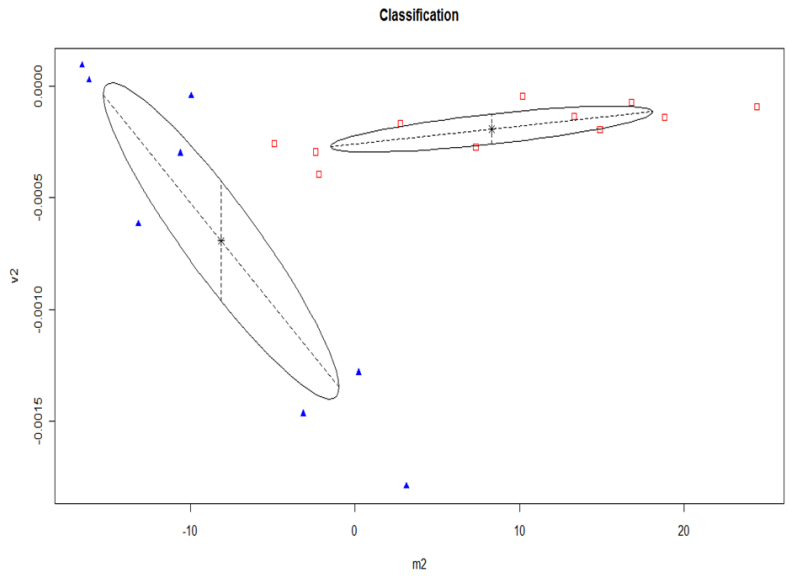
Figure 9.16: Beanplot Time Series (BTS) Model Based Clustering (3)



higher, the central or the minimum value. The results are depicted in figure 5 and figure 6. We decided to work with minima because the BIC seems to approximate better the models. By considering the data interpretation, we can observe that there are two different clusters with different characteristics in which the same results are evident. According to the previous method, the observations tend to position themselves with similar economic characteristics near to each other. In particular, we can observe the monetary areas and markets characterized by geographical regions and influential areas. In the second part of the analysis we consider different subperiods and moving windows to identify the structural changes over time. In particular the period 1-21 is relatively stable over time. In particular observation 31 and 32 show a relevant structural change due to the financial crisis.

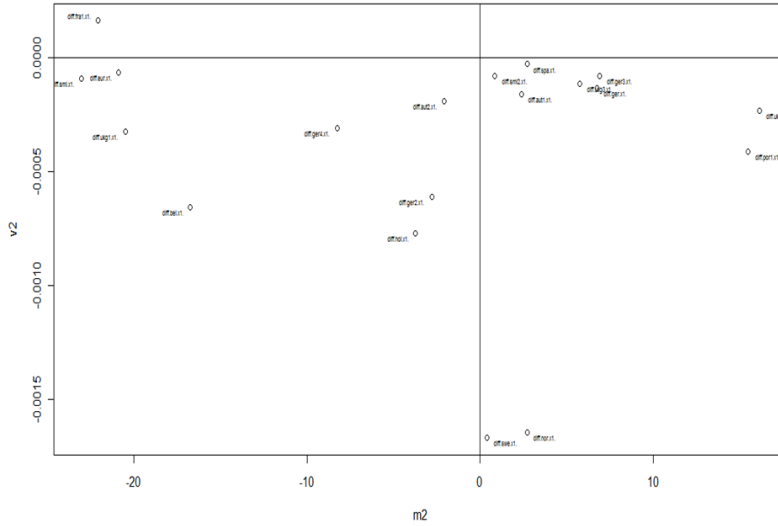
9.5. Feature Model Based Clustering for Beanplot Time Series (BTS)

Figure 9.17: Beanplot Time Series (BTS) Model Based Clustering (4)



As it is possible to observe that the contagion and the dynamics are strongly related to the different local "models" so it is also possible to observe different situations between the various observations. In practice we are able to observe the way in which the different markets tend to react differently to the financial crisis. Each market seems to be strongly related to the countries in the same system (figure 9.14 to figure 9.19), however different systems react differently. The interpretations of the identified phenomenon are various: financial linkages, domino effects or contagion mechanisms. For a similar interpretation of the contagion mechanisms and the different influential zones during the financial crisis of 2008 see Pillar et al. (2008) [563].

Figure 9.18: Beanplot Time Series (BTS) Model Based Clustering (5)



9.6 Clustering Beanplots Data Temporally with Contiguity Constraints

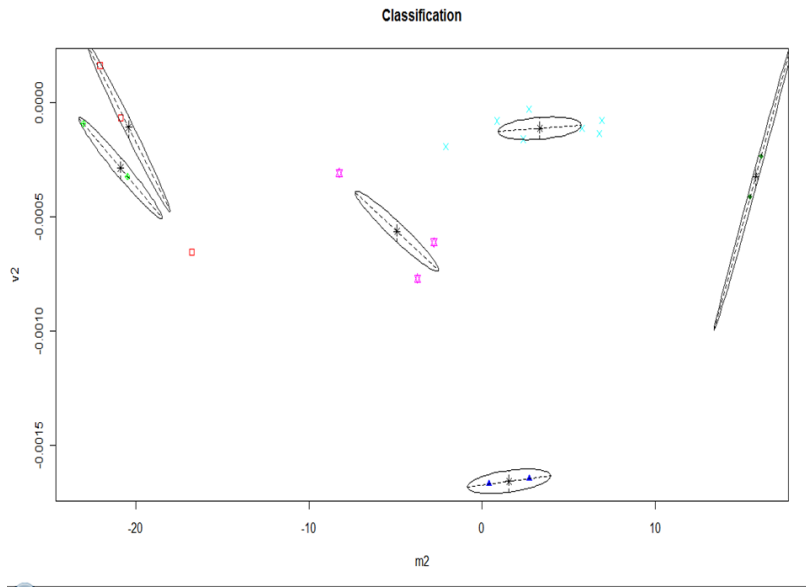
It is possible to consider the clustering of a series of beanplot symbolic data through the time with constraints of time contiguity⁴. This type of analysis could be useful, for example, in identifying similar behaviors of the beanplot time series (BTS) over time, as well as structural changes, change point, cycles and seasonalities in the intra-day dynamics. In financial analysis, for example, it could be very useful in understanding some interesting patterns in the discovered data.

In this case we are specifically considering the different groups of

⁴Murtagh 1985 [524] Gordon 1999 [320]

9.6. Clustering Beanplots Data Temporally with Contiguity Constraints

Figure 9.19: Beanplot Time Series (BTS) Model Based Clustering (6)



a beanplot time series (BTS) over time. So, with beanplots we tend to classify the subsection of the series over time to discover similar periods as internal characteristics. Contiguity constraints mean that Beanplots must be clustered in a sequential way and groups of observations need to be contiguous. At the same time this type of analysis could help to identify some outliers that need to be considered in advance before building forecasting models. In this sense, the cluster analysis can be very useful in model identification in order to identify the relevant information sets in forecasting models. The relevant sets can be the value for building forecasting models in a rolling scheme or can be used to improve the performances using a Search Algorithm (see the Forecasting Chapter).

9.7 Clustering using the Wesserstein Distance

Following Irpino and Verde 2008 ([398]), the authors propose the Wesserstein distance to cluster histograms. The same distance can be used to cluster the density data. In particular the Wesserstein $L2$ metric is proposed in Gibbs and Su (2002) [305].

Where the quantile functions of the two distributions are F_i^{-1} and F_j^{-1} so we use this distance:

$$d_W(B_{y_i}, B_{y_j}) = \sqrt{\int_0^1 F_i^{-1}(t) - F_j^{-1}(t) \, dt} \quad (9.17)$$

At the same time Irpino and Romano (2007) [395] have proved that the distance can be decomposed as:

$$d_W^2 = (\mu_i - \mu_j)^2 + (\sigma_i - \sigma_j)^2 + \sigma_i \sigma_j (1 - \rho_{QQ}(F_i, F_j)) \quad (9.18)$$

Where $\rho_{QQ}(F_i, F_j)$ can be considered the correlation of the quantiles of the two distributions F_i and F_j that could be considered in the classical QQ plot. In particular by using this distance we can classify the single density data, where we compute the dissimilarity matrix (Algorithm 16. So in that sense, we consider various different methods to classify the different beanplots over time.

Data: n Beanplot time series -BTS $\{b_{Y_t}\} t = 1...T$

Result: A vector with n elements assigning each Beanplot b_{Y_t}
for $p_1 \dots n$

begin

Choice of the I temporal interval to use

Choice of the n descriptor points to represent

Choice of the h bandwidth to use

for $t \in T$ **do**

Estimating the coefficients p

Clustering the Beanplots b_{Y_t} using the Wesserstein
Distance

end

Does the model fit adequately?

if *the model fit is not adequate* **then**

change the temporal interval I , number of descriptor
points n or the bandwidth h

end

Are the clusters fitting the data adequately?

if *the clustering method is not adequately fitting* **then**

change the number of descriptor points n or the
bandwidth h

end

end

Algorithm 16: Beanplot clustering using the Wesserstein Distance

9.8 Comparative Approaches: Clustering beanplots from Attribute Time Series

The different methods seen are related to different objectives and belong to different approaches. In the case of the coefficients estimation based on mixtures the objective is to cluster specifically the models (internal models) so the appropriate distance is the distance of Romano Lauro Giordano 1996 [594]. In general this type of coefficients estimation is related to the differences between different beanplots over time, and can be compared by considering different subperiods in the beanplot time series (BTS) where the objective is to cluster each model, separately considered, in the series of the beanplots. In that sense we are clustering the structural part of the beanplots. Here we use the coefficients estimation related to the mixtures

A completely different approach is that related to considering the trajectories of the beanplot and its synthesis. Here the representation considered is related to the coordinates. In that sense we can consider the factor time series of the beanplot time series (BTS) and we use the appropriate distance to cluster the different time series. In this case we are considering the different temporal evolution of the different series. We define this method as the "classical" one.

A modern approach is Model based clustering. In this case we are considering both the X^C the Y^C for each beanplot time series (BTS). Clearly also in this case we consider the representation by coordinates. In this case we consider also for each time the specific representation.

Finally, if the interest is to cluster each beanplot data in a beanplot time series (BTS) we can consider the different beanplot and cluster the beanplots using other types of distances such as Euclidean or the Wasserstein distance.

9.9 Building Beanplot Prototypes (BPP) using Clustering Beanplot Time Series (BTS)

A relevant Clustering application is to build prototypes (or indicators) from the original Beanplot Time Series (BTS) (Algorithm 17, Algorithm 18 and figure 9.20). In particular we start from the original Beanplot time series (BTS) which we parameterize following the two different approaches (coefficients estimation A_t and coordinates X^C and Y^C). So we obtain the Factor Time Series (defined BFT) by the Time series factor analysis methodology. In this sense we are synthesizing the information in the beanplots and we are considering the latent factors which impact on the beanplot dynamics. Then we can use a clustering technique to obtain clusters of time series and his prototypes. These prototypes are very useful to build indicators for the initial beanplot time series (BTS) related to specific groups with similar dynamics (for example starting from the Beanplot time series (BTS) related to different stocks it is possible to build indicators related to stocks with similar behaviors). At the same time the prototypes are useful in the Forecasting processes where it is possible to identify different Beanplot time series (BTS) which could be modelled to take into account the specific inter-relationships between them. Beanplot Time Series (BTS) which act as outliers can be well detected in this way.

9.10 Sensitivity and Robustness of the Clustering Methods

The outlier identification is a relevant but not simple task (see Huber 1981 [373]). So a first possible outcome in the cluster analysis is that

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1 \dots T$

Result: n Beanplot prototypes (BPP) one for each k group

begin

Choice of the I interval temporal to use

Choice of the n points to represent

Choice of the h bandwidth to use

The parameters are Beanplot model coefficients ?

if *the parameters are Beanplot model coefficients* **then**

for $t \in T$ **do**

Estimating coefficients p

Clustering the factor time series using the model distance

Deriving the Beanplot Prototypes (BPP) from the clusters

end

end

Are the clusters fitting the data adequately?

if *the clustering method is not adequately fitting* **then**

change the number of descriptor points n or the bandwidth h

end

end

Algorithm 17: Building Beanplot Prototypes (BPP) using Beanplot clustering (Model coefficients approach)

Data: n Beanplot time series (BTS) $\{b_{Y_t}\} t = 1...T$
Result: n Beanplot prototypes (BPP) one for each k group

```

begin
    Choice of the  $I$  interval temporal to use
    Choice of the  $n$  points to represent
    Choice of the  $h$  bandwidth to use
    if the parameters are Beanplot descriptors then
        for  $t \in T$  do
            Estimating the  $X^C$ 
            Estimating the  $Y^C$ 
            Estimating the factor time series (BFT) for each
            beanplot time series (BTS)  $\{b_{Y_t}\} t = 1...T$ 
            Clustering the factor time series for  $X^C$  and  $Y^C$  for
            each beanplot time series (BTS)  $\{b_{Y_t}\} t = 1...T$ 
            Deriving the Beanplot Prototypes (BPP) from the
            clusters
        end
    end

    Are the clusters fitting the data adequately?
    if the clustering method is not adequately fitting then
        change the number of descriptor points  $n$  or the
        bandwidth  $h$ 
    end
end

```

Algorithm 18: Building Beanplot Prototypes (BPP) using Beanplot clustering (Descriptor points approach)

Figure 9.20: Building Beanplot Prototypes (BPP) from the Beanplot Time Series (BTS)

$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=1	Bft=1	Prototype 1
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=2	Bft=2	
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=3		Prototype 2
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=4		
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=5		
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=6		Prototype 3
$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$	Bt=7		

of identifying the different outliers that could be specifically found in the considered data. In our case, outliers can be internal models or single beanplots, or the initial beanplot time series (BTS).

At the same time, the clustering methods are inherently based on homogeneous data with heterogeneous clusters that do not present outliers (Fritz García Escudero Iscar 2011 [283]). So a strategy that could be performed by considering the clustering methods seen before is that of starting with the complete number of observations and firstly identifying some relevant outliers. Then, there is the repetition of the clustering to identify the relevant groups.

The different clustering methods are inherently different among themselves and that need to be considered for different purposes. At the same time, the different methods exhibit different levels of ro-

bustness they need to be considered both with or without the outliers found previously in the analysis of identifying the data structures.

An alternative is the use of a robust method in clustering along one of the approaches followed in literature (see Hardin Rocke 2004 [339] García-Escudero, Gordaliza, and Matrán, C. (2003) [292] and for different approaches Henning 2009 [357] and Riani 2004 [582])

9.10.1 Ensemble Strategies in Clustering Beanplots

Another possible approach to improve the quality and the robustness of the cluster solutions in the clustering process is to adopt an ensemble strategy (see Day 1986 [171], Hornik 2005 [366], Strehl and Ghosh 2002 [643]). In practice it is possible to use different clustering methods and reconcile the information obtained by the different methods (Consensus Clustering).

In particular in the beanplot time series (BTS) it is possible to cluster by considering different methods (both clustering the time series of beanplots, or the single internal data) so we can use the ensemble methods to assess the stability of the clusters obtained.

In this sense, we can consider as input type in the ensemble clustering the following: the different internal modelling approaches, the different distances and the different methods explored above during the chapter, with the aim of comparing the results between them.

9.11 Clustering: Usefulness in Financial Applications

Clustering Beanplots time series can be useful in different contexts. For example, it is possible to monitor the behavior of different stocks

to determine the subset of the stocks determining the weight of each stock in the portfolio (for asset allocation purposes). So in this sense the characteristic of the beanplot time series (BTS) is that of using a high quantity of available information to cover different possible events in order to choose the different subset of stocks under different temporal contexts. At the same time it is possible to apply clustering techniques to evaluate the differences between stocks. It is also possible to identify the different market phases by considering a clustering technique with contiguity constraint. We obtain a contiguous series of the most similar beanplots by segmenting the initial beanplot time series (BTS) to indicate different market phases. Therefore it is possible to detect the different market phases and to determine which type of events is causing the stock behavior. The beanplot can take into account volatility and so this can be a tool for financial risk management. At the same time, the beanplot can be useful in statistical arbitrage in order to identify very similar pairs of different stocks (see Chapter 11).

Possible applications: Market Monitoring, Macroeconomic and Financial Analysis, Mining Financial Data, Statistical Arbitrage, Asset Allocation, Quantitative Trading—event identification, Event Studies, identifying patterns related to specific financial events in a temporal window, Tactical Asset Allocation—pattern identification and exploitation (using financial market inefficiencies), Risk Management—identification of the market phases.

Summary Results: Clustering
Three Clustering approaches are considered: the first one is to cluster the data model over the time using the appropriate distance.
In the second one the Beanplot Time Series (BTS) are clustered considering classical distances. By time series factorial techniques we obtain a representation of the initial Beanplot Time Series -BTS (a synthesis)
A third modern approach is based on Model Based Clustering and considers jointly all the characteristics of the Beanplot time Series (BTS).
Cluster Analysis can be used to detect outliers in the Beanplot time series (BTS) or Change Points.

Chapter 10

Beanplots Model Evaluation

In this chapter we compare the different methodologies of evaluations in clustering and forecasting beanplots as external models, and at the same time internal modelling.

We have seen in Chapters 7, 8 and 9 internal and the external modelling methods, here we will be studying the measures of evaluation of these models. At the same time, it is possible to hypothesize that there will be impacts of internal modelling on external modelling. So these methods can be considered fundamental for the correct representation of the data models or the single beanplots as internal representations. For example problems can exist in internal models as outliers, in structural changes or in specific data structures. So we need to analyse in this chapter the relations between the accuracy of the external models with respect to the Internal Modelling. At the same time there are cases in which Internal Models do not accurately represent initial data either because of the characteristics of the complex time series or because they are influenced explicitly by outliers (for the phenomenon of the overgeneralization see Chapter 1). In these cases there is a general loss of accuracy in the models. There are in this sense various strategies to consider when facing these types of problems. It is possible to

weight these observations differently or not to even consider them, or use a statistical imputation strategy etc.

A more general solution could be that of respecifying the internal models, and so attempt in this way to improve the results of the external models. Another important problem is how to identify observation in the specification phase (using some strategies like the Forward Search: in Atkinson Riani and Cerioli 2004 [47]). Different evaluations of the models can take different decisions on various data aspects, for example the interval temporal (see Chapter 6). An ineffective external or internal modelling can lead to the choice of a different data representation (as seen in Chapter 4).

10.1 Internal Modelling: Accuracy Measures

In this case, for each internal model, we need a specific strategy to evaluate the goodness of fit. In the case of the mixture models it is possible to use the chisquare as an index to measure the adequacy of the data to the model (see Du 2002 [239], and Titterington Smith and Makov 1985 [660]).

As already observed in the internal modelling part of the work we can observe that the adequacy is computed for each step of the iterative model selection in the internal modelling phase, where the goal is that of reaching the maxima data approximation of the model. The different approaches in Internal modelling evaluation are summarized in table 10.1

It is important to note that the coordinate approach for the density trace representation is related to a description of the original density data with respect to a specific modellization.

Table 10.1: Internal modelling evaluation

Approach	Evaluation Method
Mixture models	Chi square statistics
Coordinates	-

10.2 Mixture Models and Diagnostics

Mixture analysis can be used in the analysis of the internal method of the period considered (figure Algorithm 19). In particular we estimate the parameters (in particular mean, standard deviation and the proportion of the observations in the group) of two or more univariate normal distributions. The number of the distribution is given by a-priori theoretical considerations or by simple preliminary data exploration (which usually delivers some initial starting points for the algorithms). A first relevant diagnostic tool is the chi square statistic to observe the adequacy of the internal model to data (see Du [239]). The ANOVA test usually follows for each beanplot considered over time.

In particular, given a specific period of time defined by a Beanplot we can carry out a mixture analysis to study the differences between groups of observations in the defined temporal interval. To compute the mixtures it is possible to use the EM algorithm (Dempster et al. 1977 [193]), whilst in choosing the number of groups one can apply the AIC Akaike Information Criterion (Akaike 1974 [11]) with a small-sample correction (we have used also the software PAST).

$$AICc = 2k - 2\ln(L) + \frac{2k(k+1)}{n-k-1} \quad (10.1)$$

Here k is the number of the parameters, n is the number of data and L is the likelihood of the model. The lower the AIC, the better the number of mixtures that avoid the overfitting and produce the best fit ¹. This procedure compares the number of groups considered with the analysis and choosing of the optimal. So we need to compare the mean and the standard deviation for each group and minimize the AIC in the parameterization.

Each observation can be assigned to each group considering the maximum likelihood approach (see Hammer 2011 [335]). Various methods can be used in this respect as a non-hierarchical clustering method. The usefulness of the method in the diagnostic of the models is that data with poor performance with respect to one group or another one usually calls for a different coefficients estimation or descriptor points representation.

10.2.1 Application on real data: Evaluating Internal Models: the case of the Mixtures

We consider the Beanplot time series (BTS) for the Dow Jones market 1990-2010 (figure 10.1). We estimate the coefficients of the model using the mixture approach, so we obtain the coefficients π_1 , π_2 , and π_3 . At this point we can obtain at the same time the mean of the mixtures μ_1 , μ_2 and μ_3 that represents another important indicator in the adequacy of the internal models. We obtain for the first 20 periods (see table 10.2):

¹ In the computation the software Past is used: see the Past documentation [758] and Hammer 2011 [335]

Data: A Beanplot internal model $\{b_{X_t}\} t = 1...T$ in a temporal interval k

Result: n groups and an assignment of the i observations to the n groups

begin

Choice of the n groups

Assignment of the i observations to the n groups

for $t \in T$ **do**

Choice of the n groups

Assignment of the i observations to the n groups

Is the mean for the interval statistically significant?

if *the mean is different* **then**

| Use this information for the statistical arbitrage k

end

Kernel Estimation of the internal model using a different interval k

end

Is the internal model not fitting data adequately?

if *the internal model is not adequately fitted* **then**

| change the interval temporal k

end

Kernel Estimation of the internal model using the interval k

end

Algorithm 19: Mixture analysis as evaluation of the internal model

10.3 Forecasting Evaluation Methods

It is possible to evaluate the methods used in forecasting². Here, we analyse the performances of the forecasting methods by introducing some measures of adequacy of the performances obtained for the time series of attributes. In this sense, we apply these indices of adequacy for each point computed.

Definition 9. Local Forecast error E_t is the difference between the actual value of an attribute time series y_t and the predicted value of the attribute time series F_t . Here we denote for the following equations: E_t as the prediction error, F_t the value to forecast and n typically the number of the observations.

$$E_t = y_t - F_t \quad (10.2)$$

Definition 10. The forecast error for the Beanplot Time Series (BTS) of attribute (BMAE) b_{X_t} at $t = 1...T$ is:

$$BMAE_{a,t} = \sum_{a=1}^n \left(\frac{\sum_{t=1}^n |E_t|}{n} \right) \quad (10.3)$$

Where a is a single attribute of the beanplot. The index is computed as the sum of the Local Forecast error E_t considering all attributes.

At the same time it is possible to compute the Beanplot Mean Absolute Percentage Error (BMAPE), so we have:

$$BMAPE_{a,t} = \sum_{a=1}^n \left(\frac{\sum_{t=1}^n \left| \frac{E_t}{F_t} \right|}{n} \right) \quad (10.4)$$

²West 2006 [697] Hyndman 2006 [379] Hyndman Koehler (2006) [386]

In both cases the objective of using a forecast temporal interval $t = 1...T$ is to minimize the index. The indexes of adequacy are not the unique indicators of goodness of fit of an external forecast model, in fact, also a poor approximation of the internal models on a specified benchmark can be a signal of bad approximation of a forecast. Therefore, it is necessary to weight differently the beanplots that are not fitting well in the real data: an indicator of goodness of fit in the internal model, given an optimal bandwidth h obtained the Beanplot Internal Model Error for each beanplot b :

$$BIME_{b,t} = \sum_{t=1}^n h_t^* - h_t \quad (10.5)$$

If it is the case that the BIME index is performing in a poor way it means it is necessary to respecify the external model.

10.3.1 Forecasting evaluation procedure

After having considered the single forecast, the forecast error is the difference between the real value at time t and the forecast value for the correspondent period.

A Measure of error (at time t) is:

$$FE = \text{forecasting error} = 100\% \times \frac{|y_{actual} - y_{forecasting}|}{y_{actual}} \quad (10.6)$$

for the measures of aggregate error (for more than one period) each of them has different performances due to the different ways to handle outliers and observation outside some range. So we use a battery of indexes to evaluate the forecasting models. Here E_t is denoting the error, y_t denotes the series to predict and N is the number of observations. So we have:

Mean Absolute Error:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - F_t| = \frac{\sum_{t=1}^n |E_t|}{n} \quad (10.7)$$

Mean Absolute Percentage Error:

$$MAPE = \frac{\sum_{t=1}^n \left| \frac{E_t}{y_t} \right|}{n} \quad (10.8)$$

Symmetric MAPE (see Hyndman 2006 [379]):

$$sMAPE = \text{mean} \left(200 \frac{|y_t F_t|}{(y_t + F_t)} \right) \quad (10.9)$$

Percent Mean Absolute Deviation:

$$PMAD = \frac{\sum_{t=1}^n |E_t|}{\sum_{t=1}^n |y_t|} \quad (10.10)$$

Forecast Skill (related to the MSE)

$$MSE = \frac{\sum_{t=1}^n E_t^2}{n} \quad (10.11)$$

$$SS = 1 - \frac{MSE_{\text{forecast}}}{MSE_{\text{ref}}} \quad (10.12)$$

10.3.2 Discrepancy Measures

Giudici 2006 [312] shows various methods that could be used in this case. An important class of methods are related to the Discrepancy measures, for example, distances on intervals that could be used to evaluate the prediction of the size of the beanplot.

10.3.3 Applications on Real data: Evaluating the Mixture coefficients estimation and Forecasting

We consider the data for the Dow Jones Index for the period 1990-2011. We describe here the forecasting procedure:

1. Defining an adequate temporal interval (for example a year)
2. Estimating coefficients of each beanplot in the time series (the π model coefficients)
3. Obtaining the factorial time series
4. Forecasting the factorial time series (individual forecasts, combinations, or hybrid models)
5. Diagnostics of the internal and external model

We obtain the forecasting models and we can perform the diagnostic for each model.

Individual forecasting model on the factorial time series (Auto-ARIMA algorithm)

Forecast combination strategy with model selection. Weighting proportional to the forecasting performance of the models: VAR, Setar, Exponential Smoothing, Auto-Arima, Theta, Splinesf. Best methods: Auto-Arima and Splinesf with weights 0.35 and 0.65

10.3.4 Applications on Real data: Evaluating Forecasting the Dow Jones Index

Dataset 1: Dow Jones dataset 1928-2010. We consider an example related to forecasting the index for the Dow Jones (for the year 2010).

We consider for exploratory purposes the Dow Jones data (1928-10-01–2010-7-30): in total 20,549 observations. Not all the observations are used to build the models but are considered only in the visualization of the entire beanplot time series (BTS). We consider as Forecasting model period (1998-08-03–2008-08-03) so we define a first set of relevant observations in Forecasting. The objective is the Forecasting of the year 2009 and for the interval 2009–2010. Forecasting methods used: VAR, Auto-Arima, Exponential Smoothing, Smoothing Splines. At the same time as considering Forecasting combinations (Mean, Exponential Smoothing, Auto-Arima) we compare the forecasts obtained with those obtained by the naive model. At that point we consider different diagnostic measures for the models considered. The results in the table appear to be good and seem to suggest the use of forecast combinations in order to improve the results (see table 10.3 and table 10.4).

10.4 Clustering Evaluation Methods

At the same time as evaluating the Internal Models it is also possible to evaluate the External Models as Clustering and Forecasting methods. Typically the internal models can be evaluated by observing the adequacy of the beanplot model data. Here we start with the analysis of the Clustering Methods for the evaluation of the Clusters obtained. For a review of the clustering evaluation methods see Wagner and Wagner 2006 [685] and Meila M. 2003 [501].

10.4.1 Internal Criteria of cluster quality

An internal criterion can be based on the idea that a cluster that minimizes the distance within different clusters is better than one that maximizes the distance. It is an internal criteria, one based on the

10.4. Clustering Evaluation Methods

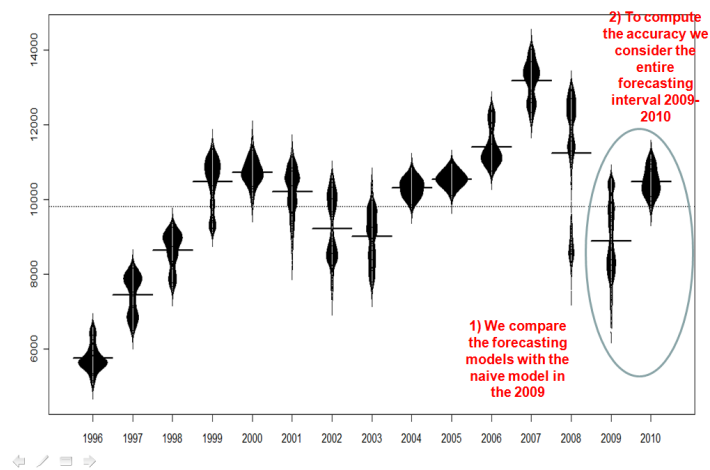


Figure 10.1: Beanplot Dow Jones Data 1996-2010 (see Drago and Scepi) 2010

results of a specific dataset and not on external information. In that sense we can consider some indices of adequacy like the Davies Bouldin and the Dunn Index . The best known are the work by Davies Bouldin 1979 [168] defined the Davies Bouldin Index (DBI):

$$DBI = \frac{1}{n} \sum_{i=1}^n \max_{i \neq k} \left(\frac{\mu_i + \mu_k}{d(ce_i, ce_k)} \right) \quad (10.13)$$

In this respect n is the number of clusters obtained by the procedure, c_x is the centroid of cluster obtained x where μ_x is the average distance of all elements in cluster x to the centroid ce_x and $d(ce_i, ce_k)$ is the distance between the centroids.

And at the same time, the Dunn Index, by Dunn 1974 [242] can be defined:

$$DI = \min_{1 \leq i \leq n} \left\{ \min_{1 \leq k \leq n, i \neq k} \left\{ \frac{d(i, k)}{\max_{1 \leq j \leq n} d(j)} \right\} \right\} \quad (10.14)$$

where $d(i, k)$ is actually measuring the distance between two generic clusters i and k . The intra-cluster distance in the cluster j , between any pair of elements is represented by $d(j)$, by using for example the maximal distance.

The inter-cluster distance $d(i, k)$ between two clusters can be represented also by any type of distance, for example a distance between the centroids of the clusters can be used.

Both these indexes aim to show the adequacy of the cluster analysis performed. It is clearly desirable that procedures perform with the lowest Davies Bouldin Index and the maximum Dunn Index possible. This result is expected for the reason that it is better that a clustering process generates clusters with high intra-cluster similarity between the components of each groups and low inter-cluster similarity.

10.4.2 External Criteria of cluster quality

In this case we consider external evaluation criteria (for example, expert evaluations), so the clustering process is evaluated by considering these criteria as the benchmark. Some examples of these indexes are the Rand and the Jaccard Index. The Rand Index, Rand 1971 [575]:

$$RI = \frac{TPV + TNG}{TPV + FPV + FNN + TNG} \quad (10.15)$$

In this case TPV and TNG represent the true answers: respectively positives and negatives. In the denominator, the sum of the total of cases where FPV is the number of false answer positives and the FNN is the number of false answer negatives.

The Jaccard Index, Jaccard 1901 [400] can be defined as the number of unique elements common to both sets (S_1 and S_2) also defined as the size of the intersection of the sets divided by the total number of unique elements in both sets (or the union). So we have:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (10.16)$$

At the same time the Fowlkes Mallows Index, by Fowlkes Mallows 1983 [273]

$$F_{a,b} = \frac{N_{1,1}}{\sqrt{(N_{1,1} + N_{0,1})(N_{1,1} + N_{1,0})}} \quad (10.17)$$

Where in the numerator appear the elements in the same cluster a and b set, and in the denominator the sum between the same cases and the others.

10.4.3 Computational Criteria

A criteria to evaluate the cluster quality is also proposed by Suzuki and Shimodaira 2004 [649], Shimodaira 2004 [628] and Shimodaira 2002 [627] which proposes an algorithm to assess the uncertainty in hierarchical cluster analysis.

In practice the P-Values of each partition are computed by bootstrap so every cluster can be evaluated. In our sense we can evaluate every cluster as an external model.

10.5 Forward Search Approaches in Model Evaluation

The Forward Search (see Atkinson Riani Cerioli 2004 [47], Atkinson Riani 2004 [46] and Riani 2004 [582]) Approach can be used to analyse the presence of outliers in the internal models but also in the external models.

In particular in the external models, the methods are applied directly on to the trajectories related to the coefficients estimation and the descriptor points

In this approach, the observations as outliers are detected, but also structural changes over time are detected so it is possible to operate to a respecification of the models.

Rolling approaches can be useful in detecting the different data structures that could be found in data, so detecting structural changes over time.

10.6 The Internal and the External Model Respecification

If the model is not adequate, then it can be respecified either by considering the outliers found or the different periods shown in the data (for example different structural changes).

Various respecification strategies can be adopted. The outliers can be imputed, respecified or not considered in a comparative approach with the initial data. A different approach is that of considering different specifications for the internal models, such as different temporal intervals, different bandwidths or different Kernels.

At the same time different respecifications can be adopted for the external models, like different periods for building external models (or the forecasting models), different methods for forecasting the attribute time series or the trajectories generated by the coefficient estimations and the descriptor points.

The process ends when there can be some satisfaction in the modelling and in the clustering or forecasting process. By citing Box and Draper's work: "essentially, all models are wrong, but some are useful" (Box and Draper 1987 [98]).

10.6.1 Application on real data: Model Diagnostics and Respecification

We start by using as a univariate forecasting model the Smoothing Splines for the three X^C attribute time series. Results are compared with the real value, where the previous observations represent the naive forecasting model. We outperform the naive model. In particular we perform an error of 5-6 with respect to the real value (MAPE 5-6). The result seems to be good. In the case of the lower part of the beanplots (the lower risk interval) this could be considered more

difficult to predict in conditions of high volatility and market crisis. So we expect in these cases a lower performance in the prediction models. By considering an alternative and competitive forecasting univariate model (by Automatic Arima: see Hyndman 2008 [382]) we outperform the naive forecasting model but we also outperform the smoothing splines approach. In particular the Automatic Arima algorithm selects the best Arima model by minimizing the AIC (Akaike Information Criteria index). So in this case the best models tend to outperform the previous ones we have considered. Here we consider the Y^C attribute time series associated to the X^C attribute time series in the stationary framework as an example of the VAR forecasting model. Here as well, we outperform the naive model, but the results are not as good as in the X case. In fact the Y contains all the complexity in our data and so the predictions are therefore necessarily less accurate (MAPE around 20). Here for predicting the Y attribute time series we use the Smoothing Splines approach in which we are specifically trying to denoise the dynamics of the shape. Two times out of three the model outperforms the naive method. The performances of the methods for the Y are not as good as the X case (and that could be expected given the nature of the attribute time series). The Smoothing Splines approach seems for the Y case the best approach in the forecasting of the Y attribute time series. Here we use the smoothing splines of the previous example while considering another relevant element: an original algorithm that optimizes the forecasting model by selecting the relevant temporal information (by minimizing the MAPE in the validation period of the model). So the procedure is divided into two distinct parts: running the algorithm to minimize the MAPE in the validation period and using the interval temporal for the forecasting. In that way we are explicitly selecting all the relevant set of information (without structural breaks) in our data. The results seem to be good (at least in respect to the other models used as benchmark). The decision to deal with a selection algorithm of the optimal

interval for the Y attribute time series can be explained by saying that we are dealing with a very volatile attribute time series (the Y) and with dynamics with frequently occurring structural changes. So we need to consider the relevant information in order to try to minimize the noise that could lead us to not correctly understand their dynamics.

See table 1 to compare the different MAPE for the Y as different attribute time series. The final conclusion of the application is that it is necessary to take into account the Search algorithm to find the best model before attempting a respecification of the models (for example by considering a different bandwidth).

Summary Results: Model Evaluation

All the models, both internal and external, need to be evaluated.

The evaluation needs to be conducted before considering the internal models, and the model that does not faithfully represent the initial data needs to be discarded.

Outliers need to be identified and eventually imputed or discarded.

At the same time, the clustering and the forecasting procedures need to be evaluated in order to improve their performances.

Bad model performances lead to model re-specification, in order to obtain better performances.

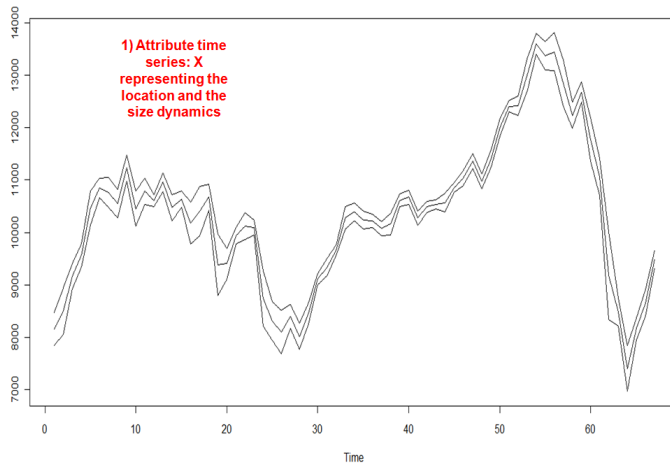


Figure 10.2: Attribute Time Series

10.6. *The Internal and the External Model Respecification*

Table 10.2: Internal modelling diagnostics

Model	π_1	π_2	π_3	μ_1	μ_2	μ_3	chisq
1	0.35	0.32	0.33	247.07	274.69	282.18	0.00
2	0.00	0.49	0.51	305.60	320.73	321.72	28.31
3	0.33	0.33	0.34	194.60	243.21	272.96	0.00
4	0.32	0.32	0.36	101.39	140.45	174.19	0.00
5	0.34	0.36	0.30	52.38	64.32	76.48	0.00
6	0.34	0.33	0.33	61.09	95.30	97.25	0.00
7	0.44	0.20	0.36	94.02	95.81	104.03	0.00
8	0.32	0.36	0.32	103.38	119.83	139.63	0.00
9	0.23	0.48	0.29	152.08	158.47	176.96	0.00
10	0.28	0.40	0.31	132.62	178.01	189.61	0.00
11	0.46	0.16	0.38	120.50	127.89	147.53	0.00
12	0.36	0.29	0.35	136.04	140.59	150.76	0.00
13	0.08	0.68	0.24	123.72	131.90	147.30	0.00
14	0.33	0.34	0.33	119.78	123.19	123.19	29.59
15	0.00	0.70	0.30	104.33	104.92	112.52	10.83
16	0.11	0.52	0.38	124.12	134.32	139.72	0.00
17	0.36	0.16	0.48	137.56	144.99	147.53	0.00
18	0.33	0.35	0.33	157.09	165.68	186.34	0.00
19	0.35	0.30	0.34	171.54	200.25	204.34	0.00
20	0.32	0.33	0.35	176.28	176.28	180.04	7.65

Table 10.3: External modelling diagnostics

	z1
ME	-0.34
RMSE	0.58
MAE	0.36
MPE	15.17
MAPE	16.98

	z2
ME	-0.18
RMSE	0.18
MAE	0.18
MPE	12.48
MAPE	12.48

Table 10.4: Accuracy of the Forecasting Models on the Attribute Time Series

Attribute time series	method	1	2	3
X	Smoothing	6.79	7	3.28
	Splines			
X	Auto	7.23	0.87	4.22
	Arima			
X	Combination	2.18	2.72	2.12
	Forecasts			
Y	Smoothing	24.11	34.92	24.54
	Splines			
	with			
	Search			

^a The considered accuracy measure in the table is the MAPE.

Chapter 11

Case Studies: Market Monitoring, Asset Allocation, Statistical Arbitrage and Risk Management

The aim of this final chapter is to show the methods seen in the previous chapters in real contexts of Finance application: Market Monitoring, Asset Allocation, Statistical Arbitrage and Risk Management. Here, an application on real time data is presented.

A second aim of this part is to show the way in which the proposed methods could improve both the classical methods based on aggregate representation (say, intervals, boxplots or histograms) and scalar data. An important characteristic of these data is that they are in real time, so the data are collected until 26/9/2011, but the models can be updated because these data are coming from different world markets, day by day. They are Indexes so they represent the compared behavior

of the world markets over time. We present some applications of the methods presented in the thesis based on real data.

11.1 Market Monitoring

Here the objective is to consider, in real time, the market changes for many different stocks by considering the beanplot time series (BTS) compared to the other types of representation. In this respect, visualization techniques are very useful and important in order to identify rapid market phases and effects caused by market shocks such as financial market news, new regulations, etc. From the visualization of the beanplots it is possible to consider clustering methods or forecasting methods to make better decisions on market operations (for example, related to risk profiles).

The original dataset of scalar data can be represented as follows. The dates are from 1/1/1990 to 26/9/2011 (figure 11.1) and they come from Yahoo Finance, in which we select the time series related to the closing prices of the most important world stockmarket indexes (Dow Jones or DJI, Dax, Cac 40 etc.). In all these cases the interesting point is that we do not observe the different time series due to the different scales and the data quantity. We need to consider some alternative representations (for example: intervals, boxplots, histograms or beanplots) for a better data visualization and exploration.

The visualization of the time series is in figure 11.2 and figure 11.3. By considering the original time series we can observe the different scales. It is clear in this case that visualization is very difficult and we cannot observe the details by considering each time series. At the same time we could be interested in considering some specific intervals (temporal intervals) to analyse the intra-period variation. This type of analysis could be very interesting in comparing the risk profiles of the time series. By considering the Scalar time series we can observe the

11.1. Market Monitoring

Figure 11.1: Scalar dataset

	auxclose	atxclose	bfxclose	omxc20coc1-e	tetpclose	rchiclose	gdaxiclose	oseaxclose	ftsemibmic-e	omxspiclose	ssmiclose
1	351.91	2451.76	289.59	3855.68	4291.53	246.78	31005	229.93	5768.7	3368.96	10729.4
2	352.68	2469.88	288.61	3863.3	4290.5	246.53	31094	231.58	5775.5	4847	3326.82
3	350.16	2446.96	286.46	3829.36	4258.24	243.81	30976	229.44	5725.5	4806	3309.04
4	353.6	289.42	3856.48	4300.94	245.33	31073	229.44	5742.4	4824.3	3317.74	10622.9
5	354.46	2475.13	291.44	3877.96	4316.4	247.28	31128	230.42	5735.2	4854.1	3304.14
6	355.13	2455.04	290.74	3877.82	4307.37	249.85	31134	233.19	5731.8	4840.7	3318.98
7	352.18	2435.34	288.63	3846.99	4258.01	246.48	30740	230.74	5713.1	4818.7	3300.4
8	350.1	2421.63	286.65	3816.14	4208.82	247.71	30645	227.09	5669.6	4783.6	3297.74
9	351.09	2415.01	288.16	3834.11	4212.14	249.95	30863	227.99	5687.8	4800.3	3270.19
10	352.73	2424.34	288.89	3854.6	4232.36	251.85	31056	229.44	5730.1	4820.8	3293.06
11	354.24	2443.14	288.92	3881.44	4245.51	254.9	31126	230.93	5761.3	4846.7	265.3
12	354.22	2452.95	288.51	3875.02	4250.71	255.63	31094	230.53	5739.4	4823.9	3315.14
13	353.94	2437.76	288.8	3869.01	4245.55	254.81	31221	230.77	5749.5	4818.3	3290.05
14	352.75	2429.17	286.43	3842.44	4220.43	251.75	31132	228.35	5728.1	4800.8	3265.14
15	353.55	2438.95	286.13	3854.19	4213.7	253.78	31190	227.91	5738.1	4803.3	3239.05
16	353.56	2444.73	285.22	3848.71	4201.89	253.61	31042	225.82	5750.5	4812.3	3235.8
17	356.36	2475.38	287.85	3882.04	4233.95	255.7	31211	228.35	5766.1	4843.2	3266.12
18	356.51	2466.49	287.85	3879.82	4214.12	255.31	31164	228.6	5768.4	4847.1	3264.67
19	358.57	2506.92	286.96	3891.4	4216.41	256.19	31129	229.81	5773.7	4853.4	3288.02
20	357.02	2492.74	289.62	3870.35	4201.81	255.27	31008	228.16	5750.7	4832.8	3279.56
21	360.42	2494.21	291.02	3913.69	4254.85	255.56	31334	229.96	5771.4	4852.3	3311.18
22	364.41	2518.97	293.92	3939.18	4279.97	258.29	31515	232.25	5797.9	4906.2	3326.83
23	365.46	2522.37	295.3	3951.72	4296.31	261.51	31563	233.51	5795.1	4916.2	3338.35
24	364.83	2527.5	295.85	3928.94	4281.64	261.4	31537	232.74	5816.9	4908.3	3332.59
25	367.32	2566.12	297.15	3958.01	4339.28	264.19	31896	234.38	5843.1	4941.5	3365.25
26	369.13	2574.2	299.19	3981.9	4366.35	267.32	32039	237.36	5870.7	4979.8	3364.49
27	369.1	2578.65	301.06	3980.77	4371.39	264.78	32176	236.87	5881.2	4995.5	3370.62
28	368.36	2580.26	301.05	3969.62	4353.15	264.87	32101	238.62	5866.5	4990.4	3347.96
29	367.48	2569.83	301.91	3970.37	4342.01	266.39	32134	234.88	5867.2	5000	3363.2
30	371.05	2587.39	3112.27	304.12	4016.75	4387.8	269.32	32328	237.8	5915.3	5044.2
31	371.8	2603.54	3101.57	304.69	4017.11	4386.4	269.39	32307	237.77	5913	5041.8

intra-temporal variations by considering short periods or short intra-temporal intervals. The best interval temporal in this case could be considered that of the year as we are interested in an analysis of the risk. So we need to consider a higher temporal interval (more or equal to a year).

At this point we can visualize the beanplot data, where some complete information on the dynamic of the original time series can be obtained (figure 11.4 to (figure 11.6)). In particular, the information related to the long run dynamics of the series (trend and cycles) is kept, in which we can check the different intra-period variability. In this case, the analysis is very useful in understanding the long run behavior of the markets and the rise of eventual speculative bubbles. In fact we find an exponential growth of the original time series (visualized by the growth of the beanlines of the different beanplots for the DJI markets), further we need to investigate the foundation of the growth to understand better if there are some speculative bubbles. The structure of the single beanplot is useful in a long run analysis

CASE STUDIES: MARKET MONITORING, ASSET ALLOCATION, STATISTICAL ARBITRAGE AND RISK MANAGEMENT

Figure 11.2: Visualization of the time series Dow Jones 1990-2011

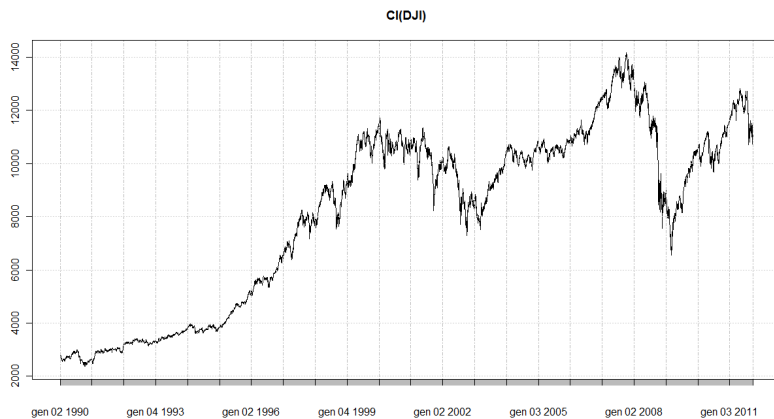
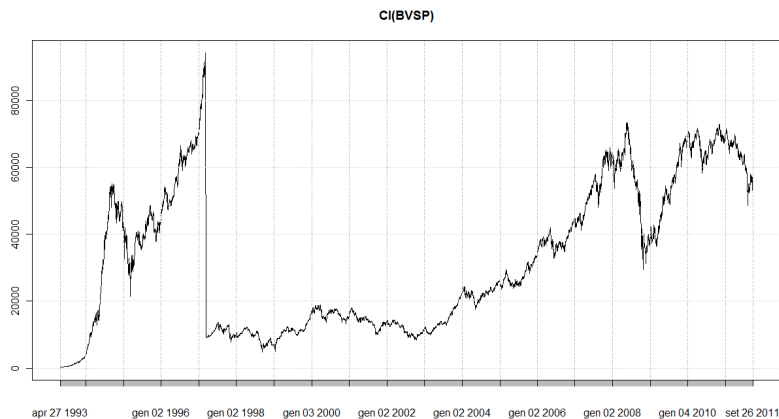


Figure 11.3: Visualization of the time series Bovespa (Brazil) BVSP 1990-2011



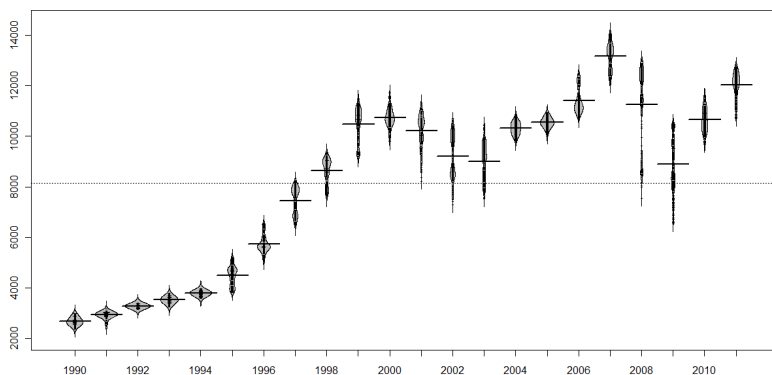
to analyze the risk profiles, where the size is higher than expected; this higher volatility is probably due to an increase in the amount

11.1. Market Monitoring

of news in the market and in structural changes due to policies, new regulations etc. The shape for each year could be considered, in order to understand over time the possible losses for the single markets, in fact by considering the time-horizon of one year we can hypothesize the possible losses year by year (before the computation of some risk measures). This could be useful to understand the financial crisis in 2008. A growth of the American economy was fuelled by various regulations and policies (as well as the growth of the dividends) and this encouraged risky behavior by some managers.

That could be visualized by considering the growth of the beanplots, as when the upper bound was reached the beanplot collapsed by creating double bumps (in 2008). Clearly the structure of the possible losses is represented by the long-run structure of the beanplot over time.

Figure 11.4: Beanplot time series (BTS) for the Dow Jones DJI 2001-2011



In the histogram time series (HTS) we can have similar information with respect to the beanplot time series (BTS), in which the informa-

CASE STUDIES: MARKET MONITORING, ASSET ALLOCATION, STATISTICAL ARBITRAGE AND RISK MANAGEMENT

Figure 11.5: Beanplot Time Series (BTS) Cac 40 (France) FCHI 1990-2011

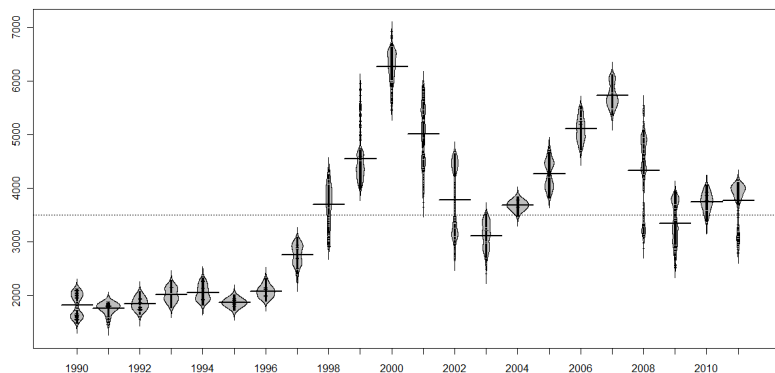
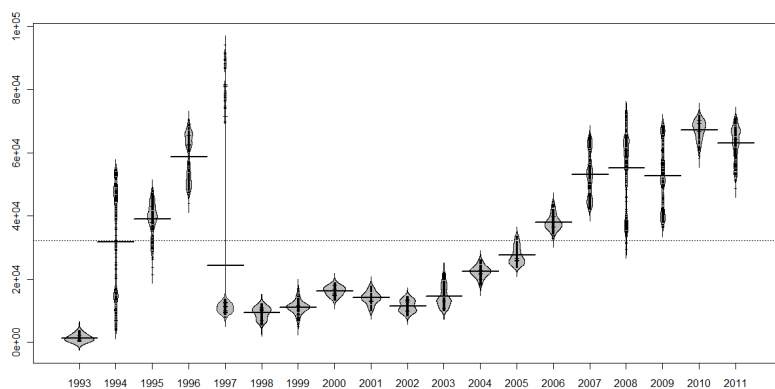


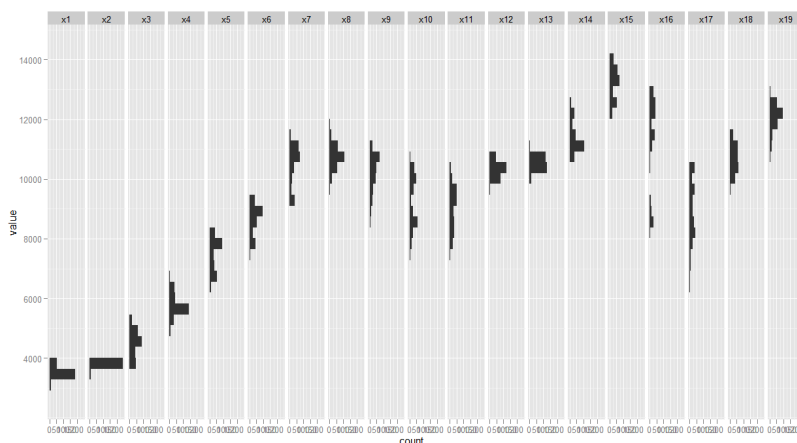
Figure 11.6: Beanplot Time Series (BTS) Bovespa (Brazil) BVSP 1990-2011



11.1. Market Monitoring

tion is more dependent on the number of bins chosen and this usually constraints the original data (figure 11.7). Various possible interpretations of the beanplots can be repeated also for the histogram time series (HTS).

Figure 11.7: Histogram Time Series (HTS) Dow Jones DJI 1990-2011



The Interval time series (ITS) shows the upper and the lower bound and the entity of the possible losses over the time. The problem for the intervals could be that we are not able to understand the difference between eventual outliers in our data. Intervals can show, as in the case of the beanplots, the risk profiles (the possible losses) in which an extreme event is represented, for example a market crash. Clearly the usefulness of the interval time series (ITS) is in the fact that it is possible to define the lower and the upper bounds in the temporal intervals. In that sense it could be useful to consider lower and upper bound in the temporal intervals, by taking into account the lower and the upper interval of variations over the time (figure 11.10).

The Boxplot time series (BoTS) tend to show a smoothed image of the intra-temporal variation over the time (figure 11.13). It is possible

CASE STUDIES: MARKET MONITORING, ASSET ALLOCATION, STATISTICAL ARBITRAGE AND RISK MANAGEMENT

Figure 11.8: Histogram Time Series (HTS) Cac 40 FCHI (France)
1990-2011

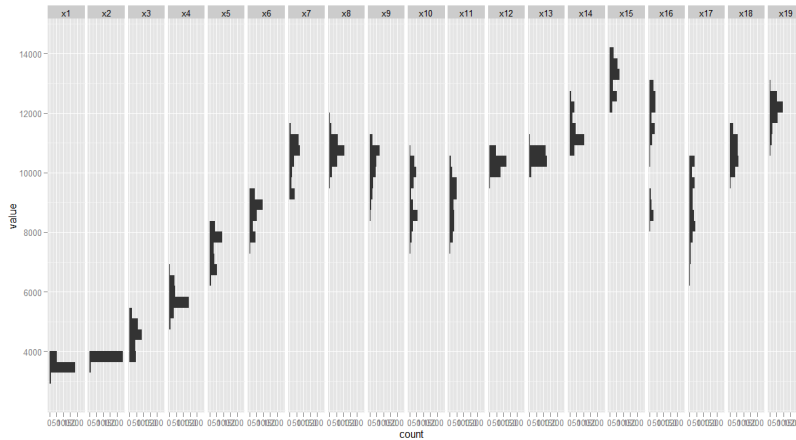
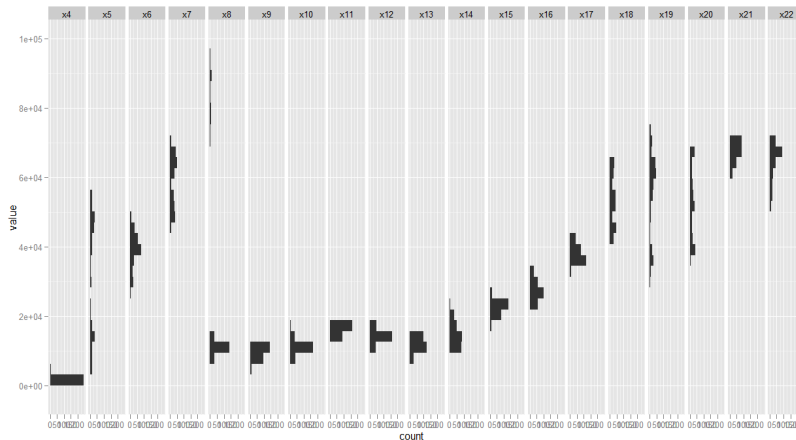


Figure 11.9: Histogram Time Series (HTS) Bovespa (Brazil) BVSP
1990-2011



11.1. Market Monitoring

Figure 11.10: Interval Time Series (ITS) Dow Jones DJI 1990-2011

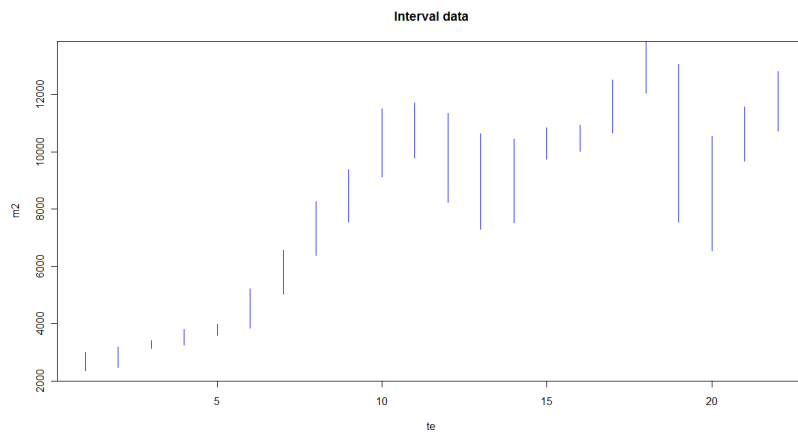
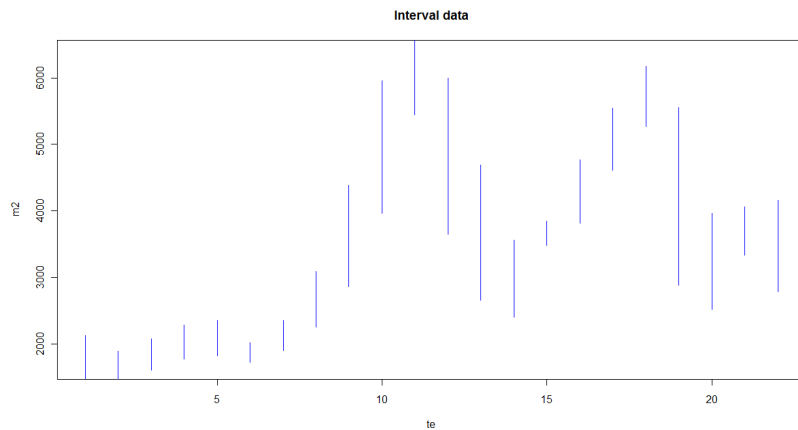
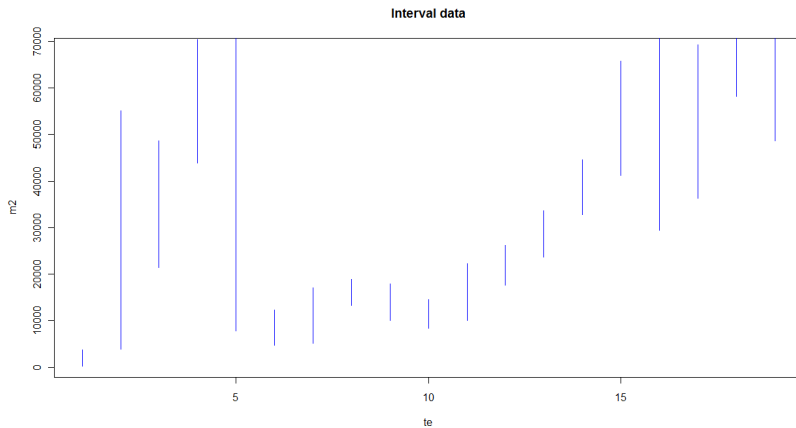


Figure 11.11: Interval Time Series (ITS) Cac 40 (France) FCHI 1990-2011



to compare the different variations over the time (and could be very useful in the analysis of the risk over the time). In any case we cannot

Figure 11.12: Interval Time Series (ITS) Bovespa (Brazil) BVSP 1990-2011



have specific information about the specific intra-temporal variation (represented by the single observations) and the different bumps.

11.2 Asset Allocation

The Asset Allocation strategies on the market can be helpful by considering both visualization techniques and also clustering. In this sense, we can identify groups of stocks with the same characteristics over time (for example, related to the intra-period variation). The quantity of data handled by the beanplot allows us to consider different market phases and the impact of various events on the time series. So, it is simpler to identify groups of similar (or dissimilar) stock for the portfolio selection.

In order to obtain the internal representations, boxplots or intervals could be used more generally than beanplots. This is due to the fact

11.2. Asset Allocation

Figure 11.13: Boxplot Time Series (BoTS) Dow Jones DJI 2001-2011

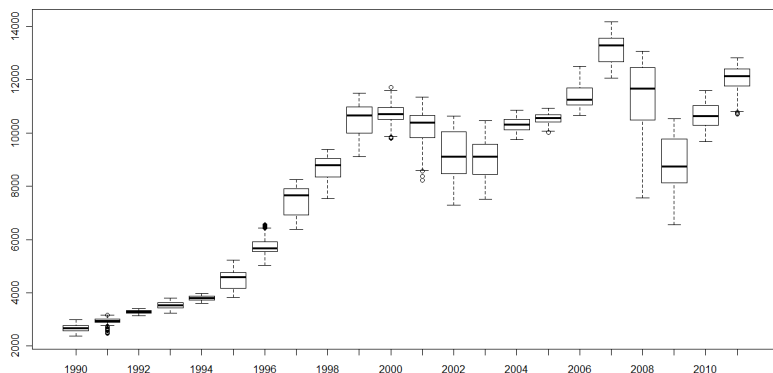
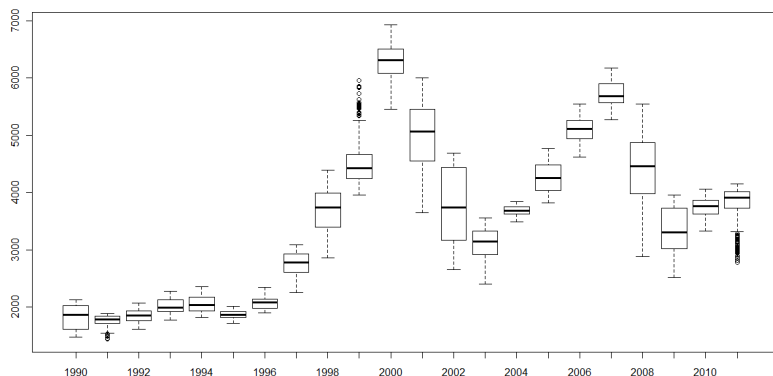
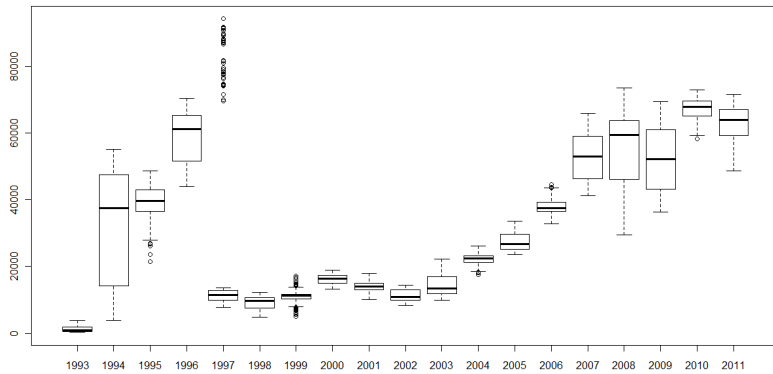


Figure 11.14: Boxplot Time Series (BoTS) Cac 40 (France) FCHI



that the densities to be computed need a specific number of observations (in the application, beanplots had no less than 36 observations per single data).

Figure 11.15: Boxplot Time Series (BoTS) Bovespa (Brazil) BVSP 2001-2011



Results seem similar for the three methods but the representations are very different. In the case of the scalars there is a problem of outliers and missing data, whereas in the aggregate representation there is the problem of the outliers (if the minimum and the maximum are considered), but not the problem of missing data. The interval tends to approximate to scalars if the temporal interval is short, then the values are similar (or the same). In this case there is no enormous difference in the trend computed for interval data and scalar data. The financial results seem good in the case of the three methods considered. The usefulness of the method of the beanplot is that we do not lose the information (as in the case of interval) when considering the same interval. The results are coherent with other types of representation. The contribution of the beanplot data is that of retaining the information of the data where there are many observations in the temporal interval. So by using these types of data we can have a twofold meaning from initial data: the relevant information of the dynamics

over the time and the structure of the intra-temporal variation.

The same observation for the intervals could not be true, infact the intervals constraints the results to the upper and the lower bound so they are useful with few observations (in short applications usually), so the results tend to be very close to those of the scalar data. Intervals do not represent correctly the intra-period variability.

The financial time series seems to be very complex so it could be necessary to take this variability into consideration in the aggregate representation (this is not possible in the intervals).

The dendrograms will be analysed and interpreted sequentially one by one: we start with figure 11.16, in which we can observe the principal influence zone at financial levels. Infact when we consider a specific shock we can observe that there are specific impacts on the financial area by its financial inter-linkages (for example, credit markets etc.). This type of analysis wants to measure the level of synchronization and interconnection between different markets and clarify the mechanisms of the diffusion of financial shocks. For example we can observe that the US market (DJI) presents characteristics similar to European markets even though maintaining its independent position. Infact we can assume that shocks on the US markets can propagate quickly on the European markets (due to the interconnections) with lags. More similar are the UK and Germany (so the markets tend to behave similarly) and France and the Netherlands or Spain and Switzerland. It is interesting to note that Japan behaves in the same way as the US: this could be considered a "signal" market that propagates the shocks that are internalized by the other markets. On the other side of the dendrogram we can observe some countries that behave differently and are clearly influenced by other shocks (for example Indonesia and Mexico). The problem of using, for example, the scalar in that case is the data imputation requested on the initial data, which could be difficult when there is more than one observation missing. Also the outliers (as in the case of the intervals) can be dangerous.

By considering the intervals in a short period the results do not appear to change using the center. So results seem to be consistent with the case of the scalars in which we avoid the problem of missing data, and the problem of outliers (in this case we need to use some different intervals that exclude outliers, for example trimmed means, etc.). In the case of the year temporal interval the results seem to be consistent with previous results, in which we can observe some short differences on the dendrograms.

By considering the Beanplot time series (BTS) and repeating the analysis above, it is interesting to note that we observe similar information for the factor F1, obtained as the synthesis of the attribute time series related to the size and the location of the beanplot time series (BTS). The F1 (see figure 11.18 and the associated dissimilarity matrix obtained in figure 11.17) shows similar information with respect to the scalar and the interval case. In this case the definition of the influence areas seems more precise. In fact there are some differences, for example Spain and Switzerland tend to behave differently with respect to the shocks. Other situations seem to be clearer: Hong Kong is very similar to Singapore, and the Japan and the US market seem to behave as "signal markets" with respect to other markets. Austria is clearly an outlier, probably due to the market size and its microstructural characteristics. Considering the second factor (figure 11.23), which is more related to the shape characteristics of the beanplot time series (BTS), we can have more information on the financial interconnections and the exact direction of the shocks on time. In particular, Mexico and Indonesia tend to have some very peculiar responses to the shocks (in fact they are particular markets characterised by very peculiar macroeconomic contexts). In the other part of the dendrograms we can observe the specific short run impacts of the shocks (that are related to the shapes of the beanplot time series BTS).

The other dendrograms are considered for robustness checks and

they confirm our initial interpretations (figure figure 11.18 – figure 11.22). In this case the results seem robust.

By considering the dendrograms with respect to the data models of the beanplot time series BTS (figure 11.27) we have some additional information, due to the latent structure of the beanplots (considered in the entire temporal interval). It is important to stress that in this case we are considering similarities between models, so the information provided is related jointly to the long run impact of the markets and the short run of markets. For example, Singapore and Hong Kong tend to be very similar over time.

Countries considered (Clustering of the models):

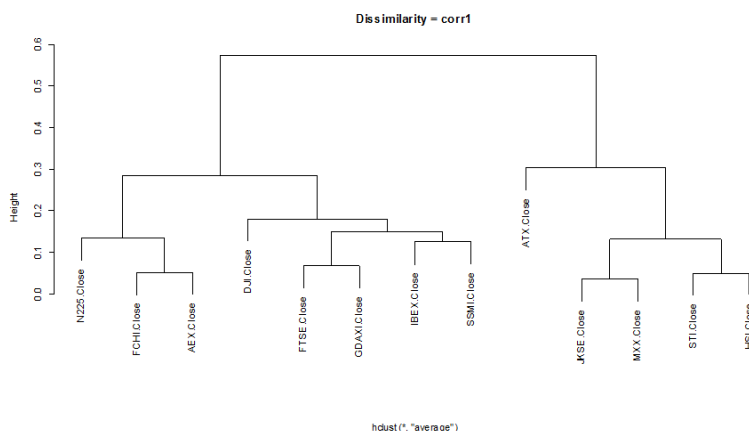
1. GDAXI
2. FTSE
3. FCHI
4. JKSE
5. HSI
6. BVSP
7. STI
8. N225
9. IBEX
10. DJI

In conclusion, it seems there are no large differences considering different interval temporal periods, but there are big differences in differentiated interval time series (ITS).

It is possible to observe that the clustering shows series with similar data structures. But the real question is: In what way is it possible to anticipate crises?

By clustering, using the model distance, we cluster the different models, in which we take into account the latent and the structural information (for each temporal interval). So we can obtain some expected results (as in the case of Hong Kong and Singapore) but also interesting results.

Figure 11.16: Clustering original time series (2000-2011)



11.3 Statistical Arbitrage

For Statistical Arbitrage, strategies identifying pairs (or groups) of stocks showing very similar characteristics over time are fundamental. By identifying some strong correlation between different stocks over time it is possible to operate and make profits on the divergences in the trajectories. Thus, the Beanplot, and the related techniques seen

11.3. Statistical Arbitrage

Figure 11.17: Clustering using coordinates correlation distance dissimilarity matrix

row.names	f.f	u.f	s.f	b.f	l.f	h.f	g.f	at.f	sp.f	sv.f	ip.f	jp.f	mx.f
1 f.f	0	0.1134360	0.6135546	0.4235974	1.044467	0.3803834	0.2429298	0.7837999	0.2858429	0.08053881	0.3235068	0.08053227	1.032272
2 u.f	0.1134360	0	0.2994363	0.3054929	0.6605234	0.4114111	0.05452794	0.5724570	0.5551930	0.0801172	0.1644887	0.2194081	0.6552998
3 s.f	0.6135546	0.2994363	0	0.0372225	0.1221333	0.1114150	0.2039112	0.1698141	0.1371788	0.3462476	0.1904640	0.5891425	0.1246917
4 b.f	0.4235974	0.3054929	0.0372225	0	0.1342986	0.1028295	0.1702793	0.2735199	0.1502713	0.3974623	0.2427422	0.6512413	0.1433974
5 l.f	1.044467	0.6605234	0.1221333	0.1342986	0	0.02187913	0.476461	0.2099025	0.4000433	0.7811704	0.4901231	1.023993	0.01233824
6 h.f	0.3803834	0.4114111	0.1114150	0.1028295	0.02187913	0	0.4250239	0.1898093	0.320713	0.4463701	0.4746347	0.944247	0.01527442
7 g.f	0.2429298	0.05452794	0.2039112	0.1702793	0.476461	0.4250239	0	0.5310863	0.1291549	0.1520152	0.2128104	0.3450733	0.4778411
8 at.f	0.7837999	0.5724570	0.1698141	0.2735199	0.2099025	0.1898093	0.5310863	0	0.2553395	0.484448	0.3291943	0.6424368	0.1614656
9 sp.f	0.2858429	0.5551930	0.1371788	0.1502713	0.4000433	0.320713	0.1291549	0.2553395	0	0.1247459	0.1958385	0.2744141	0.3584664
10 sv.f	0.08053881	0.0801172	0.3462476	0.3974623	0.7811704	0.4463701	0.1520152	0.484448	0.1247459	0	0.2104203	0.1246236	0.6968882
11 ip.f	0.3235068	0.1644887	0.1904640	0.2427422	0.4901231	0.4746347	0.2128104	0.3291943	0.1958385	0.2104203	0	0.3298147	0.4922946
12 jp.f	0.08053227	0.2194081	0.6552998	0.1246917	0.01233824	0.01527442	0.3450733	0.4778411	0.6424368	0.2744141	0.1246236	0.3298147	0.9785854
13 mx.f	1.032272	0.6552998	0.1246917	0.1433974	0.01233824	0.01527442	0.4778411	0.1614656	0.3584664	0.6968882	0.4922946	0.9785854	0
14 ol.f	0.05287241	0.2247848	0.8400297	0.8492119	1.302374	1.241056	0.3943048	1.035470	0.5143287	0.2132955	0.4501858	0.1424249	1.287287
15													
16													
17													
18													
19													
20													
21													
22													
23													
24													
25													
26													
27													
28													
29													
30													
31													
32													

Figure 11.18: Clustering using coordinates correlation distance average method

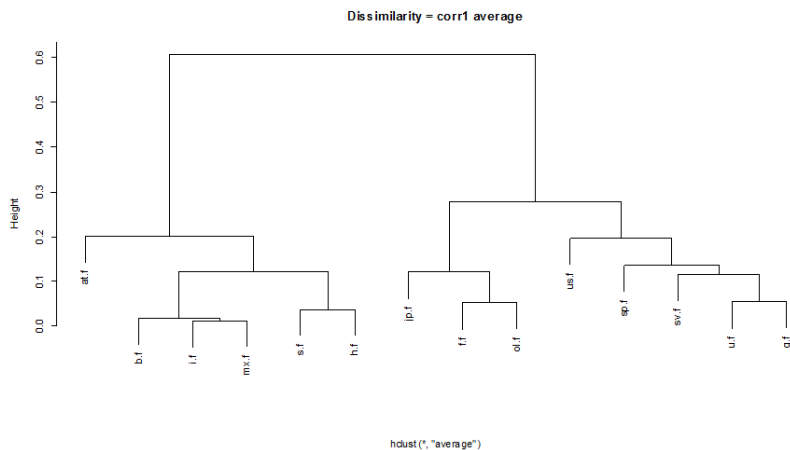


Figure 11.19: Clustering using coordinates correlation distance single method

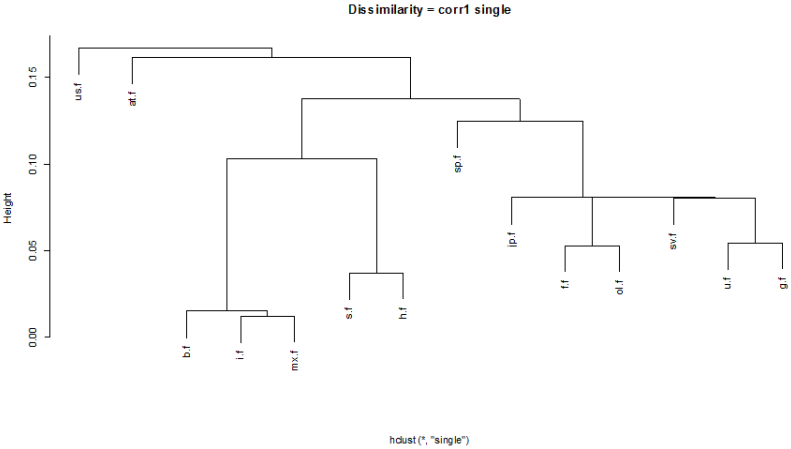
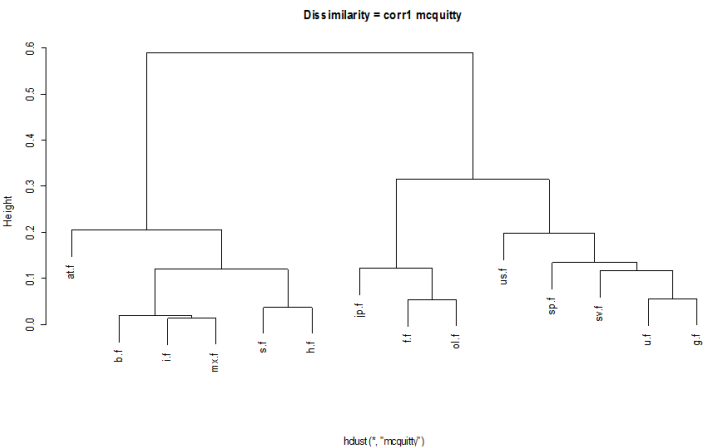


Figure 11.20: Clustering using coordinates correlation distance McQuitty method



11.3. Statistical Arbitrage

Figure 11.21: Clustering using coordinates correlation distance Complete method

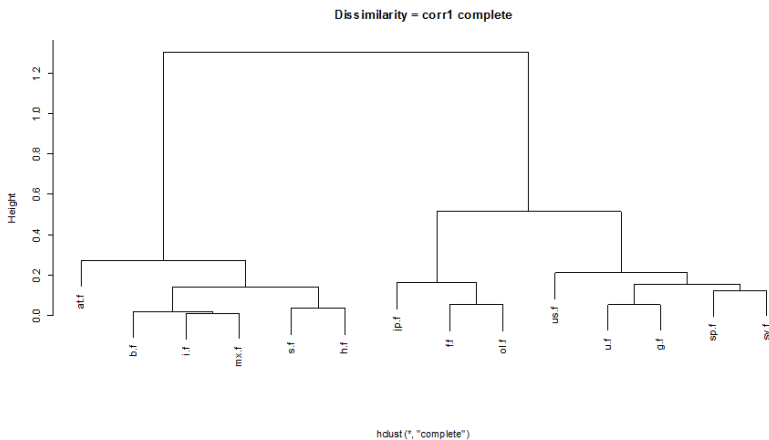
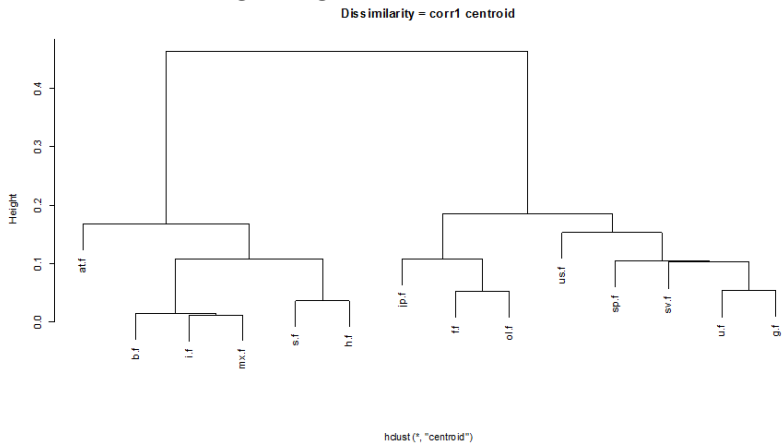


Figure 11.22: Clustering using correlation distance - Centroid method



in the thesis, helps operators in different ways, for example clustering allows the identification of some groups of stocks that can be consid-

Figure 11.23: Factor 2: correlation distance: (average)

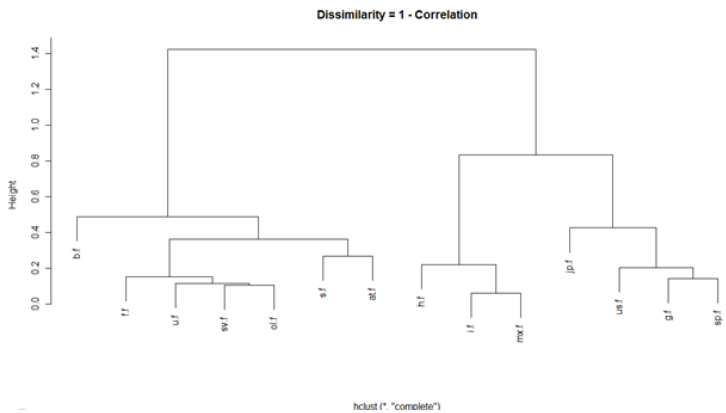
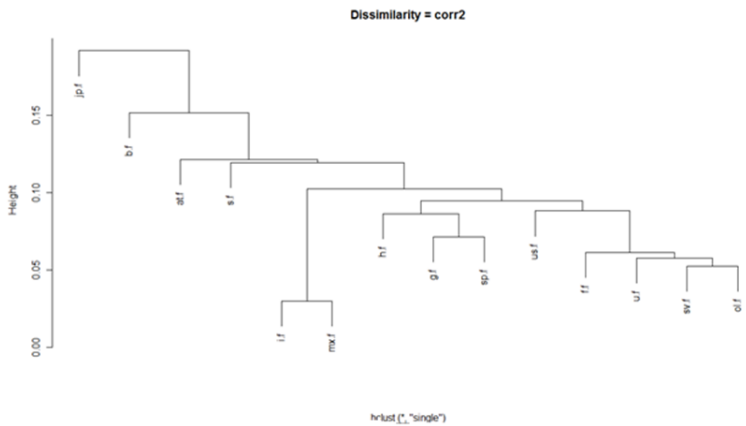


Figure 11.24: Factor 2: correlation distance: (single)



ered very similar. A subsequent statistical analysis (co-integration) can be used to build the statistical model for the arbitrage.

So we can obtain a strategy of statistical arbitrage starting from

11.3. Statistical Arbitrage

Figure 11.25: Clustering Interval Time Series (ITS) on centers: long period of interval

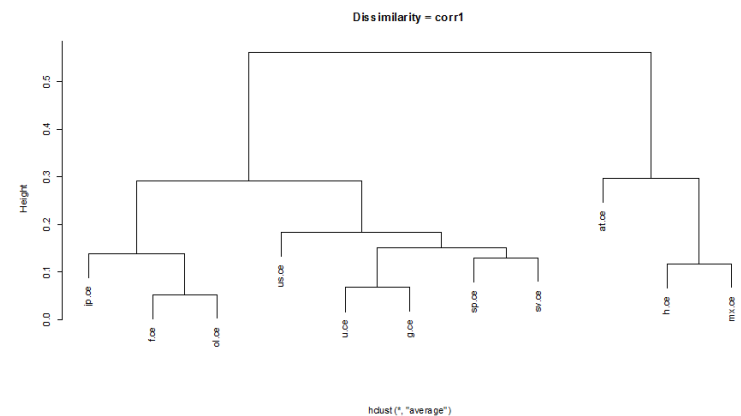


Figure 11.26: Clustering Interval Time Series (ITS) on centers: short period of interval

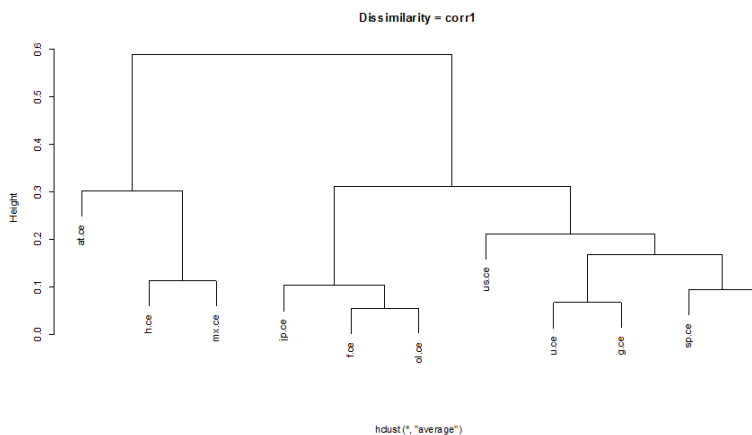
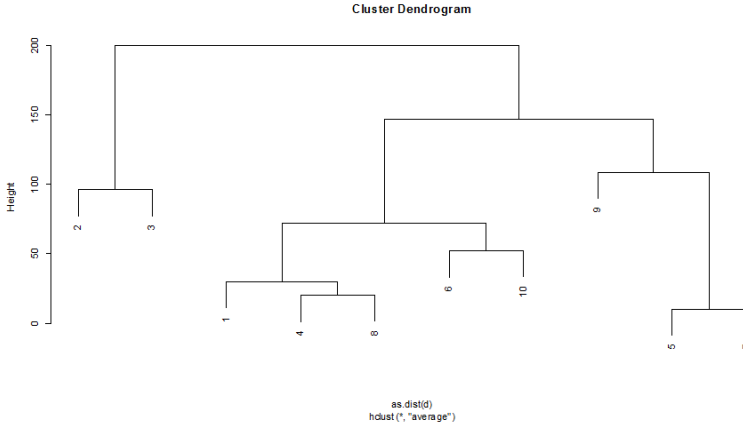


Figure 11.27: Clustering Beanplot time series (BTS) using the model distance (method average)



the clustering. From the beanplot clustering we identify the stocks that can be used in the statistical arbitrage process. We use a known statistical arbitrage model called Pair Trading or Correlation Trading. The statistical procedure is based on the Cointegration procedure in two stages of Engle-Granger¹. So we choose from the previous analysis a group of different stocks that can be related. Then we apply a cointegration analysis for the stocks together, by considering a long run model. This model represents the long run relationship between the stocks, but can be affected by a spurious regression. So we estimate:

$$\log(FR)_t = \beta + \beta_0 \log(GE)_t + \beta_1 \log(US)_t + \beta_3 \log(SP)_t + \beta_4 UK_t + \epsilon_t \quad (11.1)$$

Where FR , GE , SP and UK are the stock prices in levels for the period 01/01/2010 to 30/07/2010. We use the information collected

¹See Engle Granger 1987 [251] and Enders 1995[248]

11.3. Statistical Arbitrage

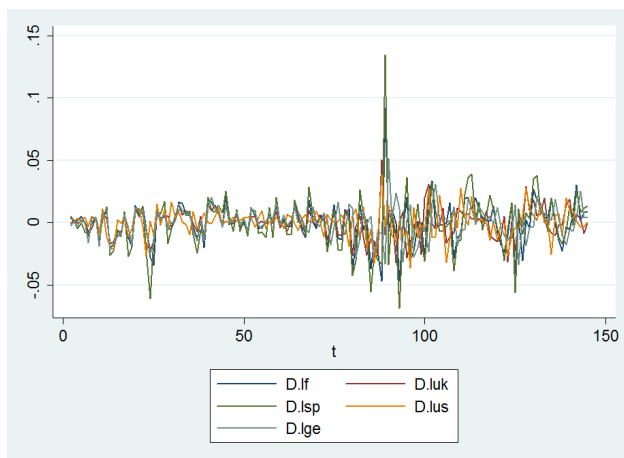
in the beanplot time series (BTS) clustering.

This model is defined as the "static" or the "long run" model because it represents the long run dynamics of the series (if the time series are cointegrated).

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.4974	0.3911	-1.27	0.2056
dat\$ge	0.1292	0.0391	3.30	0.0012
dat\$sp	0.4015	0.0238	16.85	0.0000
dat\$uk	0.5999	0.0859	6.99	0.0000
dat\$us	-0.1367	0.0985	-1.39	0.1674

Then we consider the differenced time series and we estimate a second model (the Error Correction Model).

Figure 11.28: Differenced time series from the Beanplot Clustering process



So we need to consider another model in the short run having tested

the cointegration in the model. Thus we can model the differences and the residual of the long run model.

$$\Delta \log(FR)_t = \beta + \beta_0 \Delta \log(GE)_t + \beta_1 \Delta \log(US)_t + \beta_3 \Delta \log(SP)_t + \beta_4 \Delta UK_t + z_{t-1} + \epsilon_t \quad (11.2)$$

Where the time series are differenced together and the z_{t-1} is the residual from the static model. In this sense we can model the deviations from the long run equilibrium (the static model). These deviations can be used for the statistical arbitrage using appropriate strategies.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0000	0.0006	-0.04	0.9649
dat3\$ge	0.1115	0.0470	2.37	0.0191
dat3\$sp	0.6334	0.0314	20.15	0.0000
dat3\$uk	0.0729	0.0617	1.18	0.2398
dat3\$us	-0.0950	0.0627	-1.51	0.1321
res	-0.1641	0.0471	-3.48	0.0007

11.4 Risk Management

Risk Management can benefit in various ways from these techniques. The visualization allows one to understand and compare the different risk profiles between stocks, whereas clustering helps in the identification of similar stocks in some risk profiles. The visualization and the prototypes can be used to identify early warnings for potential crises or financial problems. Using the Beanplot time series (BTS) it is possible to forecast the future levels of risk and losses over time.

11.4. Risk Management

Forecasting Beanplots means to predict the entire intra-period variation for the period considered (1 year, 1 month)- so these tools can be very useful in Risk Management Analyses.

For the scalar forecasting we consider a short period related to the first part. So we consider the period from 1/1/1990 to 26/9/2011. It is important to note that we consider the US market in the forecasting because in the previous part we observed that this market acts as a "signal" for others. So forecasting this market allows us to understand the future dynamics of the other markets. In that sense we consider various forecasting models in the period.

	Auto-Arima	ETS	Splinef	Combination
ME	-412.80	-428.64	-235.06	-358.83
RMSE	491.57	506.36	358.82	448.58
MAE	412.80	428.64	298.24	358.83
MPE	-3.79	-3.93	-2.18	-3.30
MAPE	3.79	3.93	2.74	3.30

Secondly we choose the splinef as performance and so obtain the forecasts and its confidence intervals.

	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
26	11308.67	11005.12	11612.22	10844.43	11772.91
27	11311.47	11005.18	11617.76	10843.05	11779.90
28	11314.27	11005.06	11623.49	10841.37	11787.17
29	11317.07	11004.75	11629.40	10839.42	11794.73
30	11319.87	11004.26	11635.49	10837.19	11802.56
31	11322.67	11003.60	11641.75	10834.69	11810.66

Now we will look at the interval time series (ITS). Firstly we obtain the descriptor point of the series.

Figure 11.29: Interval attribute time series DJI (first 100 observations)

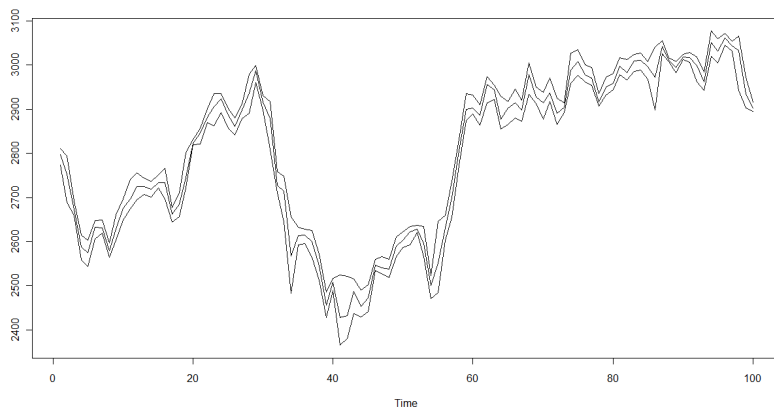
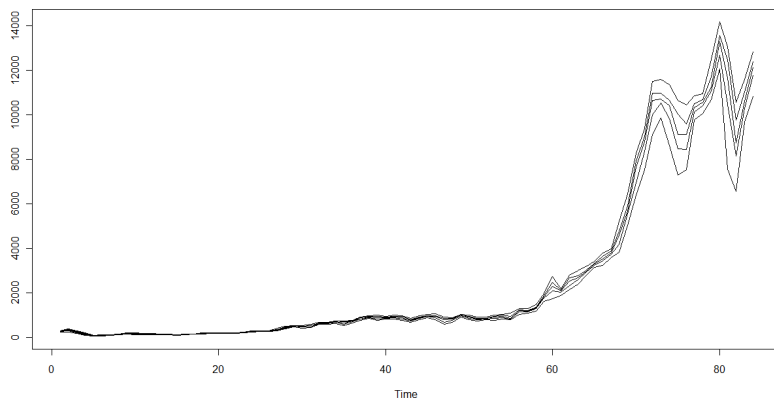


Figure 11.30: Boxplot attribute time series DJI 1900-2011



First of all we consider the series of the DJI and in particular the closing prices. We consider the beanplot time series (BTS) for the entire period. Data are related to the period 1900 to 2011 (September

2011). The first representation we choose uses coordinates. As usual we choose a bandwidth for the entire time series and we forecast the attribute time series for the beanplot time series (BTS) as well. We forecast both for the X^C and for the Y^C . The procedure is divided into three steps:

1. Forecasting the attribute time series using competitive methods
2. Forecasting the attribute time series using forecasts combinations (by an appropriate weighting scheme)
3. Detecting the best set of information using the Search Algorithm (for the Y^C)

On the X^C Forecasting we use the methods Auto-Arima and Ets ($h=3$). The result is particularly good for forecasting on the short time. If we obtain a general MAPE on 1-5 means then it signifies we can predict well the locations and the size of the beanplot over time.

On the X^C Forecasting using combinations ($h=5$) it is interesting to note that by using combinations we can use a higher horizon for the forecasting process. The MAPE in that sense is higher than in the case of the single forecasts (MAPE 7-8) but the horizon chosen is higher.

Using a Search Algorithm (on minima), the forecasting model achieves a MAPE of 1.43 as accuracy.

The aim of the Search algorithm is the identification of the best information set. In practice the analysis is divided into two steps: first, we consider the temporal intervals that minimize the error in the forecasting (for example by minimizing the MAPE), then we use this set of information in the forecasting process.

We consider the search on the Y^C , because of its volatility. The forecasting process is on $h=1$, in fact it is necessary to maximize the information and the data. In any case the search algorithm is not

strictly necessary for the X^C forecasting due to the different stability level of the attribute time series, but it is necessary for the Y^C . The Mape for the Minima (MAPE 1.43) is good after the procedure of the search algorithm.

A second step could be to Forecast using Mixtures, as we want to forecast the structural part of our beanplots over time.

The Automatic-ARIMA algorithm (and the associated ARIMA model) is the forecasting model we have chosen. By using this model we obtain a satisfying diagnostic and the forecasts obtained are (MAPE=5). So we conclude that the model can be used both for the prediction and the simulation.

The factor is well predicted by using only one method. When one method is better than others the combination is not so efficient. Using other methods it is abstractly possible to improve the results if the results of the methods are not able to discriminate the optimal one. The capability of forecasting depends on the quality of the group used in the combination.

We compare the results with the forecasting of the scalars, the boxplot, and the interval time series (ITS). The results are strictly related to the descriptors used and the data considered, so the results can vary. In any case it is important to observe that the focus of each method is very different: the scalars are a way of forecasting in the short period, so we predict observation by observation (and so we have not a feel about intra-period variability). In the case of boxplot and interval we predict using some relevant descriptors (upper and lower bound, quantiles etc.) where it is important to know that the external model adequacy depends on the general structure of the model chosen, on the complex data and its representation, and that different complex data deliver different information. In that sense the beanplot time series (BTS) are related usually to information on intra-temporal variation, that is, on more observations than in the case of interval and boxplot. So they are very useful when we need to compare long run dynamics

and wish to take into account a large quantity of information.

Figure 11.31: Forecasting Beanplot time series (BTS) using the mixture: coefficients estimation

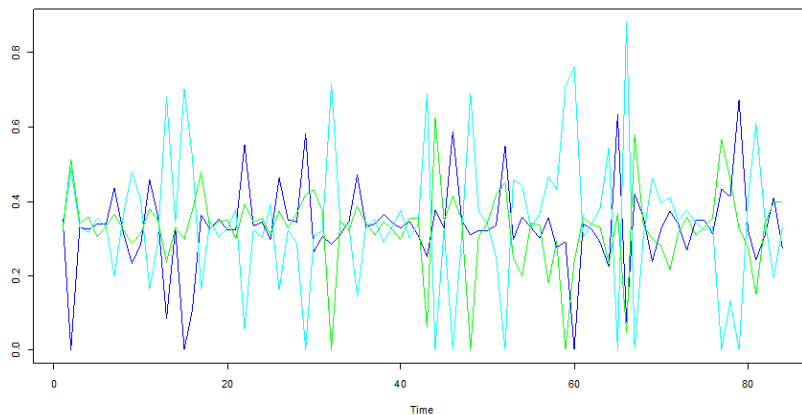
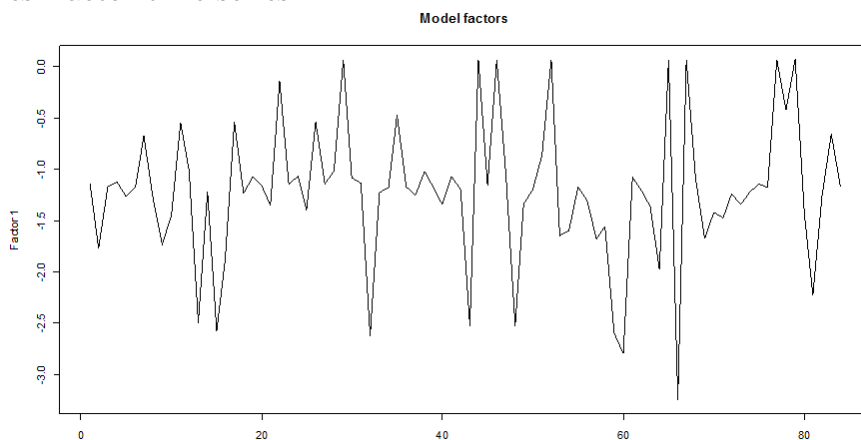


Table 11.1: Forecasting results

Forecasting	Horizon	MAPE	Descriptor	Points
Auto-Arima algorithm\ETS	3	[1-5]	X^C	
Forecasts combinations	5	[7-8]	X^C	
Splinef with Search algorithm	1	[1-34]	Y^C	

Figure 11.32: Forecasting beanplot time series (BTS) using the mix-
tures: factor time series



Summary Results: Case Studies
The algorithms presented in the work allow the replication of the methods in concrete applicative contexts, such as Statistical Arbitrage, Asset Allocation and Risk Management, for the taking of optimal decisions.
In Statistical Arbitrage (Pair Trading) it is crucial to identify stocks to use in operations such as indexes with similar characteristics. In this sense, Beanplot time series (BTS) could allow for the analysis of the long run dynamics and the selection of the most similar stocks (by beanplot clustering). Then, it could be possible to decide the trading strategy also through considering the beanplot forecasting.
In Asset Allocation strategies it is possible to decide stocks using visualization and beanplot clustering strategies
Risk Management problems can be usefully analysed by considering the beanplot visualization (which allow us to observe dynamically the risks over time due to the beanplot size and shape) and the forecasting by considering its coefficients and descriptor points.

Conclusions and Extensions for Future Research

The world of data is changing. The terms Analytics, Data Products, Data Science are frequent in the business world today. At the same time, the problem is associated with the ubiquity of huge datasets (big data) and the continuous flow of information that can add value to the business. In modern finance, developments are related to the existence of new types of data like High Frequency Data that can be considered the original data type with respect to its aggregate version, and this leads to a loss of information. In all these contexts the problem is to represent and to use adequately the information in the data.

In that sense, it is a question of using the data according to a model to gain an economic advantage on the continuous flow of information that the new technologies allow.

So the challenge for the new statisticians or the new profession, the data scientist, is not only to analyse this flow of data, but at the same time to gain knowledge and a business value from this information. In this respect, of great relevance is a sequence of well-defined phases: data collection (or data storage), data cleaning, data visualization, and finally, data analysis (or analytics) for a specific purpose: forecasting and collecting information for making better decisions. However, the main focus of the thesis is on special types of time series, defined as

high frequency data. The data are considered to be unequally spaced, to possess an overwhelming number of observations, errors (that call for special filters), missing observations, price discreteness, seasonalities and volatility clustering. These data require specific econometric methods in their analysis and the data aggregation seems not to be appropriate, due to the fact that information loss occurs.

In this thesis, we innovate this scheme by proposing a new type of data. Scalar data are used extensively in the Data Science world, and in the Financial sector. We have shown that these data sometimes represent, for example in the context of huge datasets, a difficult piece of information to use because of the difficulties in visualization (and so, in data exploration). The most frequent solution to that of using Scalar Data is that of using some types of aggregation which however result in information loss. In any case, the problem in High Frequency Data and in general in Huge Data Sets is not a loss information in itself but the representation of the underlying information. In fact the single scalar is sometimes not an adequate way to represent the data in which we want to explore some patterns of intra-data variation. This type of information could be very useful in a very important series of business or financial contexts.

We consider the changing intra-period variability or the data as genuine representations as intervals, histograms, densities or bean-plots. The intra-period variability (modelled as internal model) of the representation is important because we need to consider the pattern of variation in the time as the relevant phenomenon. If we consider the dynamics of the phenomenon over time we consider the inter-period variability. We model the intra-period variability using external models. The optimal compromise between the two models and the two representations needs to be optimized in the cycle, described in the thesis as visualization, exploration, internal modelling, clustering or forecasting (external modelling) and model evaluation. Non-optimal models can be re-specified. The actual literature does not consider

the approach to model differently from the intra-period and the inter-period variation, but this seems an important aspect of the present work because we capture the structural aspects of the data. This idea is an innovation of the present work. In particular, we are explicitly taking into account the capability of the data to capture the intra-period variability of the phenomenon (which could be used to extract value from data). At the same time the techniques required for the objectives need to be considered in real time, so throughout the thesis we consider different tools that use the changing information in a real time ideal context that both identify the relevant information and achieve the updating of the important statistical results over time.

A relevant problem running through the entire thesis is the choice of the appropriate temporal interval (for example, the choices between hour, day, week, month and year). There is no specific answer as to what is the best temporal interval in every situation. In general, the temporal interval depends on the specific application. So, in some cases it could be useful to decide a temporal interval related to short periods, for example in trading applications, whereas in risk management it is more important to consider a higher temporal interval (say, a year) to cover all the possible economic phenomena in a temporal range.

The choice of the best representation is strongly related to the choice of the best temporal interval. In particular, by choosing the temporal interval it is possible to select the best representation. Therefore, in each concrete application it is very important to define beforehand a useful temporal interval (day, month, year, etc.) then to choose the most useful internal representation in order to extract the knowledge from the data with the objective of decision making.

So for a specific reason we have worked extensively on original (very long) time series that could be considered in real time, because all our applications follow the evolution of the financial markets. We consider as well the steps of data analysis: not only considering the original time

series of the scalar data but also its aggregate representation, this maximizes the information on variability much more than the interval and the histogram. Data in this respect are obtained by a Kernel Density Estimation, in which in the final data (the beanplot data) the complete information of the location, the size (the intra-period variability) and the shape (the entire data structure) are represented.

These types of data can be particularly advantageous in the context of high frequency data. When considering an overwhelming number of observations the advantage in using the Kernel Density Estimates are clear. In other cases, we may be more interested in other phenomena, so other types of structured data can be chosen.

The visualization of these new types of data as density data or beanplot data can be considered an exploitation of all the original information available on data, because they show the initial anomalous observations of the data (the outliers), and retain the relevant information in the original data as trends, cycles and seasonalities. This information obtains new parameters (each related to a different aspect as the trend).

In the visualization part we consider for the first time the different structural changes that occur in the beanplot time series (BTS): both considering the intra-period variation (the bumps) and the inter-period variation (the change points that indicate a change in the long run dynamics). The first types of phenomena could be associated, in financial time series for example, to the arrival of specific news during the day (not so relevant), whereas the second types of phenomena could indicate a more structural variation due, for example, to the impact of new technologies and /or products on markets.

Clearly the relevant problem is in deciding which is the temporal interval to choose in the data; that is, the best type of Kernel but at the same time the bandwidth, because a different bandwidth can provide a different level of smoothness for the single beanplot data. In this respect our conclusions are that the temporal interval is crucial

because there can be relevant effects on the visualization of the cycles, and the seasonalities. The optimal beanplot data over time keep the information relevant on trend (using the beanline or the centre), the cycles and the seasonality. The beanplot visualization is the first step of an analysis and the identification of adequate models. The approach we followed is a Model Approach in which we consider the information related to the beanplot (its description over time) but we assume the beanplot to be a sum of a structural part and a residual.

In this respect it is necessary to model the beanplots in order to obtain the coefficients representing the proportions of the mixture extracted from the original data (that could be considered a sum of mixtures). The data are based on a sum of different mixtures, that is, representing the latent relevant information in data. It is important to stress the fact that in reality we consider in the data errors all inconsistencies such as missing values, errors in registrations, etc. At the same time we obtain from the beanplot another type of representation that corresponds to the fundamental description of the beanplot useful for analytical purposes. Sequentially, we have developed tools for the Clustering and Forecasting of the Beanplot Time Series (BTS). In all these cases we start from the modellization of the beanplot by using the two different strategies. The results can be different because we focus, in the Model Data of the Density, on a different temporal interval and on information which is deeper than the original data. In any case, the results of the Clustering are in line with the original data. Forecasting takes into consideration the specific identification of the models based on the Visualization, and then there is the Exploration of the structures of the descriptor point sequences (also defined as Attribute Time Series) over time.

In this sense, it is crucial to obtain a model of the factorial time series for the Data Models (representing the evolution of the structural information over time), thus allowing a forecast of the series, and also a modelling of the attribute time series for the Beanplot Data, for all

the descriptors, characteristics or attributes over time. The approach we consider is that of combining different models because of the uncertainty (and complexity) of the initial data structures. It is an observed fact that the combination of different models allows for the reduction of the uncertainty by optimizing the predictions. The results seem good when considering the fact that we are trying to capture very volatile information over time (as represented by the Beanplot shape). In this respect, as well as the use of these techniques on real data we have developed an algorithm that seeks the best set of information in the data and exploits it to obtain the best predictions over time (with respect to the occurring structural changes). The final external modelling phase is related to the Model Validation or the Model Evaluation. In this respect we compare the Clustering results with some external or internal benchmark, and more importantly we compare the results of the Forecasting with the real result in a Cross-Validation process of the Model Selection. Where we have found the best Forecasting models we can use them on real data and on real cases (for example, on a real time scenario as seen in the application to real data problems).

A clear point is related to the data analysis cycle presented during the thesis. An important element of the analysis of the internal and external models is their capability to represent correctly the original data, and its usefulness on real operations (clustering and forecasting can be considered a step for making better decisions). A relevant question to be asked by the analyst is the number of points to be considered in the internal representations. Another one could be the usefulness to consider one unique internal model, when there can be structural changes. So in that sense, change points need to be carefully considered over time. At the same time there can be cases in which data are characterised by relevant cycles or trends, so we need to consider new descriptors as the beanplot centre or its upper or lower bounds. Clearly the problem is open and solutions need to be found in the evaluation of the models and in their re-specification to obtain

even better models.

A final word about the existence of other alternatives in aggregate representations, such as Interval, Boxplots, or Histograms, might be of use. In most cases we obtain coherent information (Interval time series) for example in Clustering with Scalar time series. At the same time it is difficult to interpret the bins of the histogram.

The Beanplots allow the identification of both the Short Run and the Long Run effects on the initial data. In fact, the shape in this sense is related to the identification of the different effects over time on the groups of Beanplot Time Series (BTS). In the Forecasting the differences are structural ones. We are attempting to predict different complex objects, we are attempting to predict Interval Time Series (ITS), Boxplot Time Series (BoTS) or Histogram Time Series (HTS) and Beanplot Time Series (BTS). Each complex object can have different descriptions, but different objectives.

For example, in Interval Time Series Forecasting we are interested only in Forecasting the range between the upper and the lower bound and optionally the radii and the centre of the interval. The results can vary greatly in terms of accuracy depending on the data structure, so the methodology of the Forecasting process and its underlying assumptions are crucial in Forecasting Complex Objects.

At the same time, the Beanplots present a clear advantage in their ability to use all the information available and to reproduce the variation (or its Data Model) over time. The detection of data patterns in complex time series, for example, in financial data, which is characterized by irregular cycles, outliers, and frequent structural change, could clearly be a great advantage. The final remark is related to the results that could be obtained by using and considering this type of data or Aggregate Representations. By their nature and construction they work very well with both big data and very long time series, so they can perform well through the extraction of information from these data. For example, Risk Analysis in financial data could be improved

by the use of the data, taking into account the different structural changes that occur. So a last word is related to the use of these tools in Business Analytics and in the new field of Data Science. Those aggregate Representations that do consider an aggregation in the original data (whatever obtained) can represent a significant improvement in the process because they represent a new way to consider the Data Products, these could be not scalars but such aggregate Representations as the Beanplot (in fact, real data can be characterised by the complex volatility we are interested in representing).

The code in R (and other languages) allows us to start using these methods on real scenarios. The most relevant findings and elements of innovation in this work are in the area of visualization, internal modelling and modelling of the intra-period variation, clustering and the forecasting of these new data (the external models). What are the most relevant innovations of the present work?

The Data can be assumed to be a Mixture of Distributions. In this way we take into account the intra-temporal variation. At the same time data can be assumed to be a sum of a structural part and a noise. Data models can separate the structural part from the noise.

An Internal Representation needs to preserve complex patterns of variation intra-data. The typical intra-data representation proposed are density data (obtained using Kernel Density Estimates).

The densities show higher flexibility than histograms, in particular they tend to preserve continuity of the data (without representing them bin by bin). In obtaining these data, bandwidth is very relevant to the shape, whilst the Kernel is less used. The structural aspect and the representation of the intra-period variability by means of density data is relevant to obtain good external models (modelling the inter-period variability).

In the Visualization process, an optimal bandwidth can be obtained by the Sheather Jones method. A simulation study allows the study of the informative content of the Beanplot Time Series (BTS), with

respect to other types of Internal Representations (IR). Real Data allow a better interpretation of the Beanplot Time Series (BTS) in real contexts. In particular, we can observe the volatility levels by each day, the equilibrium levels (useful in structural changes) and the intra-period seasonalities, etc.

Two approaches are used in the Internal Modelling phase. The first, assumes the data to be a mixture and so the coefficients representing the components are considered. In this way we extract the structural information by the Beanplots. A TSFA model is used to synthesize the trajectories and so we obtain the latent factor related to the shocks changing the Beanplot structures over time. The second represents Beanplots as a whole and it uses coordinates to represent them simultaneously. In both cases, coefficients and descriptor points substitute the original data. This separation between internal modelling (to capture the internal variation) and the external modelling is relevant to obtain more information from the data, and is an important contribution of this work.

In forecasting we consider the forecasts of the TSFA model in the model-based coefficient estimation. In the second type of the approach (using the descriptor points) we forecast the attribute time series.

Various different approaches can be considered in the forecasting process but all the approaches need to be based on an identification of the external model to adopt. A combination of external models could be very useful if it is possible to find a group of forecasting models which performs well. In this case we reduce the uncertainty of choosing a unique model and we consider eventual parameter drift.

In the forecasting procedures we can use the search algorithm to improve the forecasts by choosing the optimal set of information to be included in the model. In particular, by finding the best set of information over time it is possible to obtain the best predictions, then is possible to use the optimal set of information to build forecasting models. In a second phase it is possible to apply a rolling scheme.

Various clustering approaches are considered: the first one related to the classical clustering time series. In clustering Beanplot Time Series (BTS) we have firstly considered classical distances.

Using time series factorial techniques a representation of the initial Beanplot Time Series BTS (a synthesis) is obtained. A second modern approach is based on Model Based Clustering and considers jointly all the characteristics of the Beanplot time Series (BTS). Cluster Analysis can be used to detect outliers in the Beanplot time series (BTS) or Change Points.

All the models, both internal and external, need to be evaluated. The evaluation needs to be conducted before considering the internal models, and eventually there is the discarding of the model that does not faithfully represent the initial data. At the same time, Outliers need to be identified and eventually imputed or discarded. At the same time, this aspect appears to be relevant in obtaining better prediction models. At the same time, the clustering and the forecasting procedures need to be evaluated to improve their performances. Bad model performances need to lead to model re-specification. The data cycles interrupt when the results are satisfactory both for the internal and the external modelling.

The algorithms presented in the work allow the replication of the methods in concrete applicative contexts, such as Statistical Arbitrage, Asset Allocation and Risk Management, for the taking of optimal decisions. In Statistical Arbitrage (Pair Trading) it could be crucial to identify couples or groups of stock indexes with similar characteristics. In this sense, Beanplot time series (BTS) could allow us to analyse the long run dynamics and to select the most similar stocks (by beanplot clustering). Then it could be possible to decide the trading strategy also by considering the beanplot forecasting. In Asset Allocation strategies it is possible to decide stocks using visualization and beanplot clustering strategies at the same time. Risk Management problems can be usefully analysed by considering the bean plot

visualization (which allows the dynamic observation of the risks over time due to the beanplot size and shape) and the forecasting through the consideration of its coefficients or descriptor points.

On Applications: These methods can help to identify the mechanisms of contagion between different international markets

Detection of countries that could represent a signal can be very relevant in determining an impact on other economies and in forecasting extreme values over time. It is very important that the methods proposed in the thesis are developed using R. This fact allows the use of the methods in different applicative contexts. However, for future research and possible extensions, there are cases in which it is necessary to consider more than one beanplot time series (BTS) in a modelling process. We have already considered the case of multiple time series in the case of clustering. A possible extension of the research, that of the multivariate problem related to the group of beanplot time series (BTS), is considered. In particular, it is possible to explore various topics related to the representation of the co-integration between two time series using the beanplot as a graphical tool, the time series factor analysis on the beanplot time series (BTS), the regression, and simultaneous equation modelling using recursive systems.

We can have two or more time series and test them to observe the co-integration, then we can represent the co-integration vector as a beanplot. In particular, we can consider here two or a number of scalar time series. We can consider the Engle Granger test and represent the co-integration vector as a specific beanplot over time. However, the research on this point is open, in fact it could be very important for market monitoring (or for handling very numerous data) to consider more than one beanplot time series (BTS). At the same time, it is possible to consider different co-integration procedures. The importance of this point cannot be underestimated because it is very important to model groups of beanplot time series (BTS) as well.

By starting from a group of different time series we need to model

the "market evolution". At the same time we need to represent the general dynamics of a market by considering a large number of beanplot time series (BTS). For example, we can use for each attribute time series for the descriptor points the TSFA methodology to synthesize the series as a whole. In practice, the analysis follows two distinct steps: a first step is related to the representation of the beanplot time series (BTS) using the coordinates, obtaining specifically the attribute time series; in a second step, we use the TSFA Time Series Factor Analysis or the DFA Dynamic Factor Analysis for each sequence of values and so we obtain the synthesis of the market. In all these cases various approaches could be considered, for example working on the Beanplot coefficients or descriptor points and obtaining from these the dynamics of the entire market.

A multiple time series is a different time series observed simultaneously in various contexts. Multivariate time series are time series related to a synthesis of a phenomena. There are cases in which forecasting using only one single series is not useful and it is better to predict one indicator. In general, multivariate time series could be useful for various aims: for example to measure latent phenomena and robustify the analyses.

It is possible to consider Regression Models in the Beanplot Context. Useful models in this way can be generalized from Intervals and Histogram data to Beanplots. In this sense, we can explicitly consider distances used in literature. In particular, for each method considered we can explore the use of different distances.

With the aim of analysing the dynamics of more than one Beanplot time series (BTS), we can consider a simultaneous systems of equations (SEM) of density data. The models to be implemented by considering the beanplot time series (BTS) can be predictive, in the sense it can predict the future evolution of the beanplot structure over time. So these models can allow both the prediction of the future paths of the series and the prediction of the risk dynamics (as represented by the

shape and the size in the beanplots).

So, these simultaneous models of beanplot time series (BTS) need to be focused on the forecasting and the simulation of the phenomena, just as the scalar simultaneous equation models do. For example, the simultaneous equation models in econometrics used for business cycle analysis, the forecasting of the main economic variables, and the simulation of economic policy are all useful models that represent all the phases of building a macroeconomic model for forecasting. So they can become relevant tools in risk analysis. At the same time it is possible to consider approaches to the PLS (Partial Least Squares) using Beanplots.

In all these approaches the emphasis is that of working on the forecasting by considering models with more factors determining the dynamics of the time series. In these cases we need to model not only one specific Beanplot time series (BTS) at a time but we need to model groups of time series together. Attention must be paid to modelling the beanplot time series (BTS) in order to understand the dynamics of the internal variation of an entire system, for example an economy.

These models can be very rich in their possible interpretations. Considering the internal variations for more than one time series, we can model both the aspects related to the mean (i.e. predict the effects of the shocks) as well as the effects of the shocks on the internal variability of the series (considered together). So the future challenge is that of extracting from the original huge data predictions for the future by considering entire models of more than one Beanplot time series (BTS). These models can be used in a very wide number of ways: for example, in Risk Analysis related systems, for the simulation of the analysis of different scenarios and lastly, they can be used in a general sense for the taking of better decisions.

Extensions for the future could be the application of the thesis methods in fields and operations of Finance. In particular, possible developments could be made in Risk Management, Statistical Arbi-

trage, Asset Allocation and Market Monitoring.

Another useful application for Economic and Financial analysis are the Control Charts. In practice, control charts are particularly helpful in the detection of processes that could be defined as out of statistical control and therefore in need of very attentive monitoring. So in this case, the control charts are a very useful tool for the monitoring of markets and of financial processes with the aim of Early Warning Systems.

Appendix A

Routines in R Language

The computer code in R used throughout the thesis is available upon request.

My contacts are:

Carlo Drago:

e.mail (personal): c.drago@mclink.it

e.mail (at University of Naples): carlo.drago@unina.it

website: <http://web.mclink.it/MD3687/>

Appendix B

Symbols and Acronyms used in the Thesis

B.0.1 Symbols

x_i Scalar data

y_t, x_t Scalar time series (Homogeneous time series)

y_t^f, x_t^f Scalar time series at frequency f (Homogeneous time series)

$(x_i)_i^N, (z_i)_i^N$ High Frequency Time Series (Inhomogeneous time series)

$duration_i$ Duration between two events

$(tr_{1i}, tr_{2i}, tr_{3i})$ time of the trade, as the price, volume.

$(qu_{1j}, qu_{2j}, qu_{3j})$: qu_{1j} time, bid price, ask price of the trade

$H_{n,m}$ Data matrix or Data table

$[x_i, \overline{x_i}]$ Interval data with lower and upper bound

$[x_{2i}, \overline{x}_i]_t$ Interval time series, with lower and upper bound

p_t^f time series of prices at a frequency f

r_t^f time series of returns at a frequency f

$r_{q,t}^f$ time series of portfolio q returns at a frequency f

$[x_i, \tilde{x}_i, \overline{x}_i]$ triplex data

$[m_u, q_u, Me_u, Q_u, M_u]$ boxplot data

$(I_{i,h}, \pi_{i,h})$ with $h = 1, \dots, n$ histogram data

B_t beanplot time series

b_{Y_t} beanplot data derived by the time series Y_t

$K()$ Kernel density estimation: kernel

h Kernel density estimation: bandwidth

$[a_t]$ beanplot attribute time series

$[a_{L_t}, a_{U_t}]$ beanplot lower and upper bound

$[a_{M_t}]$ beanplot beanline attribute time series

$[a_{C_t}]$ beanplot center attribute time series

$[a_{R_t}]$ beanplot radius attribute time series

$[a_{OP_t}]$ beanplot first observation time series

$[a_{CL_t}]$ beanplot last observation time series

$A_t = [p_{1,t}, p_{j,t}, \dots, p_{k,t}]'$ Internal model coefficients

I_t a measure of goodness of fit for the internal model

X^C beanplot description: coordinates (related to the x)

Y^C beanplot description: coordinates (related to the y)

F_t forecast at time t

$f^1, f^2 \dots f^m$ forecasts obtained in a combination scheme

F_t^{CM} forecasts combination

B.0.2 Acronyms and Abbreviations

SDA Symbolic Data Analysis

TAQ Trade and Quotes Database

MIDAS Mixed Data Sampling Regression Models

IR Internal (or Intra-Period) Representations

STS Scalar Time Series

ITS Interval Time Series

BoTS Boxplot Time Series

CTS Candlestick Time Series

HTS Histogram Time Series

BTS Beanplot Time Series

BFT Beanplot Factorial Time Series

DFA Dynamic Factor Analysis

TSFA Time Series Factor Analysis

BPP Beanplot Prototypes

FFT Fast Fourier Transform

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

DBI Davies Bouldin Index

RI Rand Index

splinef Cubic Smoothing Splines method

auto-arima Automatic Arima method

ETS Exponential Smoothing

MSE Mean square error

GA Genetic algorithm

KNN K-Nearest Neighbor

VAR Vector Autoregressive Models

VECM Vector Error Correction Models

SETAR Self-Exciting Threshold AutoRegressive models

MAPE Mean absolute percentage error

sMAPE Symmetric mean absolute percentage error

MSE Mean Square Error

RMSE Root Mean Square Error

MPE Mean Percentage Error

Bibliography

- [1] A.A.V.V. (2006) *Knowledge Extraction and Modeling - Workshop Introduction* September, 4th – 6th 2006 Villa Orlandi Island of Capri, Italy
- [2] A.A.V.V. *Data, data everywhere*
http://www.economist.com/node/15557443?story_id=15557443 Web. 5 Aug 2011
- [3] A.A.V.V. *Big Data, Big Problems: The Trouble With Storage Overload*
<http://gizmodo.com/5495601/big-data-big-problems-the-trouble-with-storage-overload>
 Web. 5 August 2011
- [4] A.A.V.V. *Beyond the PC* <http://www.economist.com/node/21531109>
 Web 14 October 2011
- [5] A.A.V.V. *High frequency data analysis* Published in Automated Trader Magazine Issue 05 April 2007 : Strategies, Web 23 October 2011
<http://www.automatedtrader.net/articles/strategies/600/high-frequency-data-analysis>
- [6] Adler D. (2005). vioplot: Violin Plot. R package version 0.2, URL
<http://CRAN.R-project.org/package=vioplot>
- [7] Aggarwal Charu C.(2007) *Data Streams: Models and Algorithms*, Advances in Database Systems Series Vol. 31, Springer.

- [8] Ahlberg J.H., Nilson E.N. Walsh J.L. (1967) *The theory of splines and their applications* Mathematics in Science and Engineering, New York: Academic Press.
- [9] Ahmad R. (1975) *A Distribution Free Interval Mathematics Analysis of Probability Density Functions* in Interval Mathematics (Proceedings of the International Symposium Karlsruhe, West Germany, May 20-24-1975, K. Nickel, ed., Springer Verlag Berlin, 1975.
- [10] Ahmed M., Anwei C., Xiaowei D., Yunjiang J., and Yunting S. (2009). Statistical Arbitrage in High Frequency Trading Based on Limit Order Book Dynamics. *Order A Journal On The Theory Of Ordered Sets And Its Applications* p.1-26.
- [11] Akaike H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6): 716723.
- [12] Alefeld, G. and Herzerberger, J. (1983). *Introduction to Interval computation* Academic Press, New York.
- [13] Alefeld G., Mayer G. (2000) Interval analysis: theory and applications, *Journal of computation and Applied Mathematics* 121, 421-464
- [14] Alizadeh, S., Brandt, M.W. and Diebold, F.X., (2002) Range-based estimation of stochastic volatility models. *Journal of Finance* 57, pp. 1047–1091.
- [15] Allison, P. D. (2001) *Missing Data* Thousand Oaks, CA: Sage Publications.
- [16] Andersen, T. G., (2000) Some Reflections on Analysis of High-frequency Data. *Journal of Business and Economic Statistics* 18, 146–153

- [17] Andersen, T. and Bollerslev, T. (1994). *Intraday seasonality and volatility persistence in financial markets* Working Paper No. 193, Kellogg Graduate School of Management, Northwestern University.
- [18] Andrawis, Robert R, Amir F Atiya, and Hisham El-Shishiny (2010) Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting* In Press, no. December 2008: 1-34.
- [19] Andre A. P. Nogales F.J., Ruiz E. (2009) *Comparing univariate and multivariate models to forecast portfolio value-at-risk*, Statistics and Econometrics Working Papers ws097222, Universidad Carlos III, Departamento de Estadística y Econometría.
- [20] Andreou, E., Eric Ghysels and Kourtellos A. (2010) Regression Models With Mixed Sampling Frequencies, *Journal of Econometrics*
- [21] Andreou, E., Ghysels E. and Kourtellos A. (2010) *Forecasting with mixed-frequency data* Chapter prepared for Oxford Handbook on Economic Forecasting edited by Michael P. Clements and David F. Hendry
- [22] Antoch, J., Brzezina, M., and Miele, R. (2010). A note on variability of interval data. *Computational Statistics* 25:143 (153)
- [23] Antunes C.M. and Oliveira A.L. (2001). *Temporal Data Mining : an overview* KDD Workshop on Temporal Data Mining: 1-15.
- [24] Apostolatos et al. (1968) *The Algorithmic Language Triplex Algol 60* Numerische Mathematik 11, 1968
- [25] Araújo, T. and Louçã, F. (2008) The seismography of crashes in financial markets *Physics Letters A*, 372,4, 429–434, 2008, Elsevier

- [26] Arlot S. Celisse A. (2010) "A survey of cross-validation procedures for model selection" *Statistics Surveys* Vol. 4 (2010) 4079 ISSN: 1935-7516 DOI: 10.1214/09-SS054
- [27] Armstrong J.S. (1984) *Forecasting by Extrapolation: Conclusions from 25 Years of Research* Interfaces, 14 (Nov.-Dec.), 52-66
- [28] Armstrong, J S. (2005). The International Journal of Applied Forecasting Ways to Improve Forecast Accuracy. *Foresight* 1, no. 1: 29-35.
- [29] Armstrong, J.S., 2001. *Combining forecasts*. In: Armstrong, J.S.(Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers, Boston, pp. 417-440
- [30] Armstrong, J. Scott (ed.) (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, Massachusetts: Kluwer Academic Publishers
- [31] Armstrong, J. S., Adya, M. and Collopy, F. (2001), *Combined forecasts*, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Publishers, Norwell, MA.
- [32] Arroyo J. *Symbolic Time Series Forecasting* (2009) Wienerwaldhof Workshop in Symbolic Data Analysis 2009
- [33] Arroyo, J. (2010): Forecasting candlesticks time series with locally weighted learning methods. In: H. Locarek-June, C. Weihs (Eds.): *Classification as a Tool for Research*. Springer, DOI 10.1007/978-3-642-10745-0 66.
- [34] Arroyo, J. (2011) *Some Advances in Symbolic Time Series Forecasting* Workshop in Symbolic Data Analysis Namur, Belgium, June 2011

- [35] Arroyo J. Bomze I. (2010) Shooting Arrows in the Stock Market. Compstat 2010 Paris.
- [36] Arroyo J., Espínola R. and Carlos Maté, (2011). Different Approaches to Forecast Interval Time Series: A Comparison in Finance. *Computational Economics* Springer, vol. 37(2), pages 169-191, February.
- [37] Arroyo J., Muñoz San Roque A., Maté C., and Sarabía A. (2007) *Exponential smoothing methods for interval time series* Working Paper Conference: ESTSP 07: First European Symposium on Time Series Prediction (TSP). Publication: Proceedings. City: Otaniemi, Espoo (Finland). Date: 7- 9, February, 2007.
- [38] Arroyo, J., González-Rivera, G., Maté, C. and San Roque, A. M., (2011) *Smoothing methods for histogram-valued time series: An application to value-at-risk* Statistical Analysis and Data Mining, n/a. doi: 10.1002/sam.10114
- [39] Arroyo J. Gonzáles-Rivera G. Maté C. Muñoz San Roque A. (2010) *Smoothing Methods for Histogram-valued Time Series An Application to Value-at-Risk* Working Paper.
- [40] Arroyo J., Gonzales Rivera G., and Maté C. (2009) *Forecasting with Interval and Histogram Data: Some Financial Applications*. Working Paper
- [41] Arroyo J., González-Rivera G., Maté C. (2010) Forecasting with Interval and Histogram Data: Some Financial Applications, in *Handbook of empirical economics and finance* Editors Aman Ullah; David E. A. Giles. Ed. Chapman & Hall CRC Press. 2010.
- [42] Arroyo J., Maté C., Muñoz A., (2006) *Hierarchical clustering for boxplot variables*, International Federation of Classification Soci-

eties 2006 Conference: Data Science and Classification. Ljubljana, Slovenia, 25-29 July 2006

- [43] Arroyo J. and Maté C. (2006) *Introducing Interval Time Series: Accuracy Measures* in COMPSTAT 2006, *Proceedings in Computational Statistics*, Heidelberg, pp. 1139-1146. Physica-Verlag.
- [44] Arroyo J., Maté C. (2009) *Forecasting Histogram Time Series with K-Nearest Neighbours Methods* International Journal of Forecasting, 25, pp.192-207
- [45] Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* 16, 521–530.
- [46] Atkinson, A. C.; Riani, M. (2004) The forward search and data visualization. *Comput. Stat.* 19, No. 1, 29-54.
- [47] Atkinson A.C. Riani M. Cerioli A. (2004) *Exploring Multivariate Data with the Forward Search* New York: Springer – Verlag.
- [48] Atkinson A.C., Riani M., Cerioli A. (2004) *Exploring multivariate data with the forward search*, Springer Series in Statistics, Springer, New York, 2004, 621 pages, ISBN: 0-387-40852-5
- [49] Atkinson, A. C. and Riani, M. and Cerioli, A. (2010) The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, 39 (2). pp. 117-134. ISSN 1226-3192
- [50] Avellaneda, M., and J-H. Lee (2010), Statistical Arbitrage in the U.S. Equities Market, *Quantitative Finance* 10, 761-782.
- [51] Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution, *Biometrika* 83 (4), pp.715-726.

- [52] Babcock C. (2006) *Data, Data, Everywhere* Information Week, 9 January 2006 <http://www.informationweek.com/news/175801775> Web 5. August 2011.
- [53] Bagnai A. (2001) *Un modello econometrico per lo studio dell'economia italiana* manoscritto Università di Roma "La Sapienza"
- [54] Baker Kearfott R. Kreinovich V. (1996) *Applications of interval computations* Springer
- [55] Baldini P., Figini S. Giudici P. (2006) *Nonparametric approaches for e-learning data* Working Paper
- [56] Balzanella A., Irpino A., Verde R. (2010) *Dimensionality reduction techniques for streaming time series: a new symbolic approach* In: Classification as a Tool for Research' – Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V., Dresden, March 13-18, 2009 Studies in Classification, Data Analysis, and Knowledge Organization Springer, Berlin – Heidelberg – New York (2010).
- [57] Balzanella A., Romano E., Verde R. (2008) *Summarizing streaming data via a functional data approach*, In Proceedings of the First Joint Meeting of the Societ  Francophone de Classification and Classification and data analysis group of the Italian Statistical Society, Caserta, 11-13 June 2008, Springer, Berlin Heidelberg New York
- [58] Banfield J.D., Raftery A.E. (1993). *Model-based Gaussian and Non-Gaussian Clustering* Biometrics, 49, 803-821.
- [59] Bao, Y., Lee, T.-H. & Saltoğlu, B., 2007. Comparing density forecast models. *Journal of Forecasting*, 26(3), p.203-225.

- [60] Barnett W.A. Salmon M. Kirman A. (eds.)(1996) *Nonlinear Dynamics and Economics: Proceedings of the Tenth International Symposium in Economic Theory and Econometrics (International Symposia in Economic Theory and Econometrics)* Cambridge University Press.
- [61] Barucci E. Renó R. (2002) Value at Risk with High Frequency Data. In *New Trends in Banking Management* (2002), ed. Physica-Verlag, pp. 223–232
- [62] Basalto N., De Carlo F. (2006). *Clustering Financial Time Series* in *Practical Fruits of Econophysics* 2006, 4., 252-256
- [63] Batarseh O.G. and Wang Y. (2008) *Reliable simulation with input uncertainties using an interval-based approach* Proc. 2008 Winter Simulation Conference, Miami, Florida
- [64] Bates, J. M. and Granger, C. W. J. (1969) The combination of forecasts. *Op. Res. Quart.*, 20, 451-468.
- [65] Batini C. (2010) *Data quality and data architecture as the two pillars of data governance for data analysis* SAS Institute 6 Ottobre 2010
- [66] Battaglia F. (2007) *Metodi di Previsione Statistica* Springer
- [67] Bauwens L., Hautsch N. (2006) *Modelling Financial High Frequency Data Using Point Processes* Working Paper
- [68] Bauwens, L., Hautsch, N. (2008): *Modelling Financial High-Frequency Data Using Point Processes* forthcoming in the *Handbook of Financial Time Series Econometrics*, Ed. T. A. Andersen, R. A. Davis, J.-P. Kreiss, T. Mikosch.

- [69] Beck N. 2004 *Longitudinal (Panel and Time Series Cross-Section) Data* Lecture Notes
- [70] Beck N. (2006) *Time Series Cross Section Methods* Working paper
- [71] Beck, J. B., Kreinovich, V., and Wu, B. (2004). Interval-valued and fuzzy-valued random variables: From computing sample variances to computing sample covariances. In Lopez-Diaz, M., Angeles Gil, M., Grzegorzewski, P., Hryniewicz, O., and Lawry, J., editors, *Soft Methodology and Random Information Systems*, pages 641–648. Springer, Heidelberg, Berlin.
- [72] Beckers, S. (1983) Variance of Security Price Return Based on High, Low and Closing Prices *Journal of Business* 56, 97-112.
- [73] Benjamini, Y. (1988): *Opening the Box of the box plot*, The American Statistician, 42, 257-262.
- [74] Benzecri J.P. et al. (1973) *L'Analyse de Données*, Dunod, Paris
- [75] Berkhin P. (2006) *Survey of Clustering Data Mining Techniques*. Working Paper
- [76] Berthold M.R. Hand D.J. (eds.) (2003) *Intelligent Data Analysis: an Introduction* 2nd rev. and ext. ed. 2003.
- [77] Bertrand F., Goupil F. (2000) *Descriptive Statistics for Symbolic Data*, In *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.H. Bock and E. Diday), Springer Verlag, Berlin, (2000), 103 – 124
- [78] Batarseh O.G. and Wang Y. (2008) Reliable simulation with input uncertainties using an interval-based approach. Proc. 2008 Winter Simulation Conference, Miami, Florida

- [79] Biais. B., Glosten L., Spatt C. (2005) Market microstructure: A survey of microfoundations, empirical results and policy implications, *Journal of Financial Markets*, 8 (2005), pp. 217-264.
- [80] Bifet A. Kirkby R. (2009) *Data Stream Mining: A Practical Approach*, COSI Center for the Open Software Innovation
- [81] Billard L. (2006) *Some Analyses of Interval Data*, Journal of Computing and Information Technology, 16, 4, 225 – 233.
- [82] Billard L. (2010) *Statistical Approaches for Complex Data* 19th International Conference on Computational Statistics Paris – France, August 22–27
- [83] Billard L., Diday E. (2000) *Regression analysis for interval-valued data* In: Data Analysis, Classification and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00), Springer, Belgium, pp. 369-374.
- [84] Billard L., Diday E. (2002): *Symbolic regression analysis* In: Jaguga, K. et al. (eds) Classification, Clustering and Data Analysis. Springer-Verlag, Berlin (2002)
- [85] Billard L. Diday E. (2003) *From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis* Journal of the American Statistical Association, 98, 470–487.
- [86] Billard, L., Diday, E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics.
- [87] Billard L. Diday E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley & Sons

- [88] Billard L. Diday E. (2010) *Symbolic Data Analysis: Definitions and Examples*, Working Paper
- [89] Bloom N. (2009) *The Impact of Uncertainty Shocks* Econometrica, Econometric Society, vol. 77(3), pages 623-685, 05.
- [90] Bloom N. (2011) *Shock, Paura e Recessione* La Voce Web. 14 August 2011 <http://www.lavoce.info/articoli/pagina1002493.html>
- [91] Bock, H. H. (1996) *Probabilistic Models in Cluster Analysis* Computational Statistics and Data Analysis, Volume 23, Number 1, 15 November 1996, pp. 5-28(24)
- [92] Bock, H.H. (1998) *Probabilistic approaches in Cluster Analysis* Bulletin of the International Statistical Institute 57, 603-606
- [93] Bock, H.H. (2008) *Probabilistic Modeling for Symbolic Data* in Brito P. (Editor) Compstat 2008, Physica-Verlag HD
- [94] Bock H.H., The classical data situation. In *Symbolic Official Data Analysis*, Boch H.H. and Diday E. (Eds). Springer, 24-38.
- [95] Bock H-H. and Diday E. (2000) *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg
- [96] Box, G.; Jenkins, G. (1976), *Time series analysis: forecasting and control*, rev. ed., Oakland, California: Holden-Day
- [97] Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *JRSS B* 26 211-246.
- [98] Box, G.; Draper N.R. (1987). *Empirical Model-Building and Response Surfaces*, p. 424, Wiley.

- [99] Box G. and Jenkins, G. (1970) *Time series analysis: Forecasting and control*, San Francisco: Holden-Day
- [100] Bolasco S. (1999) *Analisi Multidimensionale dei Dati. Metodi, strategie e criteri d'interpretazione* Carocci editore
- [101] Boller R.A. and Braun S.A. and Miles J. and Laidlaw D.H. (2010) *Application of Uncertainty Visualization Methods to Meteorological Trajectories*". In *Earth Science Informatics* vol. 3, no. 1-2, pp. 119–126.
- [102] Bollerslev, T. (1986) *Generalized Autoregressive Conditional Heteroskedasticity* Journal of Econometrics, 31(3), 307327.
- [103] Borman, S. (2009) *The Expectation Maximization Algorithm A short tutorial* Working Paper
- [104] Bouchaud J.P. Mezard M. Potters M. (2002) Statistical properties of stock order books: empirical results and models, *Quantitative Finance*, 2 (2002), pp. 251256.
- [105] Boukhelifa N. and Duke D.J. (2009) *Uncertainty Visualization - Why Might it Fail?*. In CHI '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pp. 4051–4056
- [106] Boukezzoula R., Galicheta S. Bissieriera A. (2011) A MidpointRadius approach to regression with interval data. *International Journal of Approximate Reasoning* International Journal of Approximate Reasoning Article in Press doi:10.1016/j.ijar.2011.07.002
- [107] Boyd D. and Crawford K. (2011) *Six Provocations for Big Data* September 21, 2011. Available at SSRN: <http://ssrn.com/abstract=1926431>

- [108] Brandes O. et. al. (1968) The Time Domain and the Frequency Domain in Time Series Analysis *The Swedish Journal of Economics* Vol. 70, No. 1 (Mar., 1968), pp. 25-42
- [109] Breitung J. Eickmeier S. (2005) *Dynamic Factor Models* Deutsche Bundesbank Working Paper
- [110] Brito P. , (1991) *Analyse de donnees symboliques: Pyramides d'heritage*. Thèse de doctorat, Université Paris IX Dauphine.
- [111] Brito P., (2007) *Modelling and analysing interval data* In Proceedings of the 30th Annual Conference of GfKl, pp. 197208. Springer.
- [112] Brito, P. and Duarte Silva, A.P. (2011). *Modelling interval data with Normal and Skew-Normal distributions*. Journal of Applied Statistics, (in press).
- [113] Brockwell P.J. and Davis R.A. (2002) *Introduction to Time Series and Forecasting*. New York: Springer.
- [114] Brown, R.L., J. Durbin and J.M. Evans (1975) *Techniques for Testing the Constancy of Regression Relationships over Time* Journal of the Royal Statistical Society, Series B, 35, 149-192
- [115] Brownlees C.T., Gallo G.M. (2006) *Financial econometric analysis at ultra-high frequency: Data handling concerns* Computational Statistics & Data Analysis, volume 51,4,2232-2245.
- [116] Buckley J.J. (2004) *Fuzzy Statistics* Springer
- [117] Burkill J. C. (1924) *Functions of Intervals* Proceedings of the London Mathematical Society 22, 375-446.

- [118] Campbell J.Y., Lo A.W., MacKinlay C., Lo A.Y. (1996) *The Econometrics of Financial Markets* Princeton University Press
- [119] Canova F. (2007) *Applied Methods for Macroeconomic Research* Princeton University Press
- [120] Carmona R. (2004) *Statistical analysis of financial data in S-PLUS* Springer
- [121] Carreras, C. and Hermenegildo M. (2000) Grid-based histogram arithmetic for the probabilistic analysis of functions *Abstraction, Reformulation, and Approximation* 107–123, Springer
- [122] Castle, J. L., Fawcett, N. W. P., Hendry, D. F. and Reade, J. J. (2007) *Forecasting, Structural Breaks and Non-linearities*, mimeo, University of Oxford, Oxford.
- [123] Caussinus, H. (1986) Models and Uses of Principal Component Analysis, in *Multidimensional Data Analysis*, J. De Leeuw et al. eds., DSWO Press, Leiden, 149-170.
- [124] Cazes P., Chouakria A., Diday E. et Schektman Y. (1997) Extension de l'analyse en composante principales á des données de type intervalle, *Rev. Statistique Appliquée*, Vol. XLV Num. 3 pag. 5-24, Francia.
- [125] Celeux G. Govaert G. (1995) *Gaussian Parsimonious Clustering Models* Pattern Recognition 28, 781-793
- [126] Cerchiello P., Figini S., Giudici P. (2008) Data Mining models for business and industry. In *Statistical practice in business and industry* (S. Coleman et al ed.), pp 163-209, Wiley

- [127] Cerioli A. Ingrassia S. Corbellini A. (2004) *Classificazione simbolica di dati funzionali: un'applicazione al monitoraggio ambientale* in Data Mining e Analisi Simbolica a cura di Davino C. Lauro C. Franco Angeli, 2004
- [128] Chalabi Y. Würtz D. (2009) *Econometrics and Practice: Mind the gap!* 25 April 2009 R/Finance 2009
- [129] Chambers, J. M. and Hastie, T. J. (1992) Statistical Models in S, Wadsworth & Brooks/Cole.
- [130] Chang, I.H., Tiao, G.C. and C. Chen (1988). Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*, 30, 193-204.
- [131] Chen C. and Liu L.M. (1993) Forecasting time series with outliers, *Journal of Forecasting*, 12, 1, 13–35, 1993
- [132] Cherednichenko S. (2005) *Outlier Detection in Clustering* University of Joensuu, Gennary 2005.
- [133] Cheung Y.L. Cheung, Y. W. Wan A.T.K., (2007) .An empirical model of daily highs and lows. *International Journal of Finance and Economics* 12, 120.
- [134] Chiu, S.-T., 1991. Bandwidth Selection for Kernel Density Estimation. *Annals of Statistics*, 19(4), p.1883-1905. Available at: <http://projecteuclid.org/euclid.aos/1176348376>.
- [135] Chou, R.Y. (2005). Forecasting financial volatilities with extreme values: the conditional autoregressive range (CARR) model. *Journal of Money, Credit & Banking* 37, pp. 561-582
- [136] Chow, G.C. (1960) *Tests of Equality between Sets of Coefficients in Two Linear Regressions* *Econometrica*, 52, 211-22.

- [137] Chatfield C., 1988. What is the best method of forecasting? *Journal of Applied Statistics* 15, 19-39
- [138] Chatfield C., 1996. Model Uncertainty and Forecast accuracy. *Journal of Forecasting* 15, 495-508.
- [139] Chiu, Shean-Tsong. 1991. Bandwidth Selection for Kernel Density Estimation. *Annals of Statistics* 19, no. 4: 1883-1905. <http://projecteuclid.org/euclid.aos/1176348376>.
- [140] Chow G.C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* 28 (3): 591-605.
- [141] Chouakria A. (1998) *Extension des méthodes d'analyse factorielle à des données de type intervalle* Thèse de doctorat, Université Paris IX Dauphine, 1998.
- [142] Chu C. S. J., Stinchcombe M., and White H. (1996) Monitoring structural change. *Econometrica*, 64 (5):1045–1065, 1996.
- [143] Cipolletta I. (1992) *Congiuntura Economica e Previsione* Il Mulino
- [144] Cipriani, M. and Guarino, A. (2005) Noise Trading in a Laboratory Financial Market: A Maximum Likelihood Approach, *Journal of the European Economic Association* April-May, 3(2-3), pp. 315-321.
- [145] Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, Vol. 5, pp. 559-583.
- [146] Cleveland, W. S. (1993) *Visualizing Data*. Hobart Press, Summit, New Jersey.

- [147] Coffman K.G. Odlyzko A.M. (2011) *Growth of the Internet*. In Optical Fiber Telecommunications IV B: Systems and Impairments, I. P. Kaminow and T. Li, eds. Academic Press, 2002, pp. 17–56
- [148] Cohen, D. J., and Cohen, J. (2006), "The Sectioned Density Plot", *The American Statistician*, 60, 167174.
- [149] Cole, A. J. and Morrison, R. (1982), *Triplex: A system for interval arithmetic*. Software: Practice and Experience, 12: 341350
- [150] Colombo, A. and Jaarsma R. (1980). *A powerful numerical method to combine random variables* IEEE Transactions on Reliability 29 (2), 126129.
- [151] Cont R. (1999) *Statistical Properties of Financial Time Series* Lectures presented at the Symposium on Mathematical Finance, Fudan University Shanghai, 10-24 August 1999
- [152] Cont R. (2001) *Empirical properties of asset returns: stylized facts and statistical issues* Quantitative Finance Volume 1 223-236 Institute of Physics Publishing.
- [153] Cont R. Long range dependence in financial markets. *Fractals in Engineering*, 159–180, Springer
- [154] Cont, R. (2011) *Statistical Modeling of High Frequency Financial Data: Facts, Models and Challenges* (March 1, 2011). Available at SSRN: <http://ssrn.com/abstract=1748022>
- [155] Cook, R. D. (1999). Graphical detection of regression outliers and mixtures. In Proceedings of the International Statistical Institute 1999 , Finland. ISI

- [156] Coppi, R., Gil, M. A. & Kiers, Henk A.L., (2006) The fuzzy approach to statistical analysis *Computational Statistics & Data Analysis*, Elsevier, vol. 51(1), pages 1–14, November.
- [157] Coroneo L. Veredas D. (2006) *Intradaily seasonality of returns distribution A Quantile Regression approach and Intradaily VaR* Working Paper
- [158] Corrales D. Rodríguez O. 2011 *INTERSTATIS: The STATIS method for interval valued data* Working paper
- [159] Corsi, F., Dacorogna, M., Müller, U. and Zumbach, G. (2001). Consistent High-Precision Volatility from High-Frequency-Data. Internal Paper, Olsen & Associates, Zürich, Switzerland
- [160] Cryer J. Chan K.S. (2010) *Time Series Analysis with Applications in R* (third edition) Springer
- [161] Dacorogna M.M- et. al. (1990). "Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis" *Journal of Banking Finance* Elsevier, vol. 14(6), pages 1189-1208, December.
- [162] Dacorogna M.M. et al. (1993) A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance*, Elsevier, vol. 12(4), pages 413-438, August.
- [163] Dacorogna M.M. et. al. (2001) *High Frequency Finance* Academic Press.
- [164] Dacorogna M.M. Muller U.A. Pictet O.V. De Vries C.G. (2001) Extremal Forex Returns in Extremely Large Data Sets. *Extremes* 4:2, 105-127, 2001 Kluwer

- [165] Daley, D.J, Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes* Springer, New York.
- [166] Dalgaard P. (2002) *Introductory Statistics with R* Springer Verlag
- [167] Dave, R.N. and Krishnapuram, R., (1997). Robust clustering methods: a unified view. *IEEE Transactions on Fuzzy Systems*, 5(2), p.270-293
- [168] Davies, D. L., Bouldin, D. W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intelligence* 1, 224-227.
- [169] Davies P.L., Gather U, Nordman D, Weinert H. (2007) *Constructing a regular histogram: a comparison of methods* Working Paper
- [170] Davino C. Lauro N.C. (a cura di) (2004) *Data Mining e Analisi Simbolica* Franco Angeli editore
- [171] Day W.H.E (1986) Foreword: Comparison and consensus of classifications. *Journal of Classification*, 3:183–185, 1986.
- [172] De Beer C.F., Swanepoel J.W.H. (1999) Simple and effective number-of-bins circumference selectors for a histogram. *Statistics And Computing* 9(1): 27–35
- [173] De Boor C. (1978) *A Practical Guide to Splines* Springer-Verlag, New York.
- [174] De Carvalho F. (1995) Histograms in symbolic data analysis *Ann. Oper. Res.* v55. 229-322
- [175] De Carvalho Francisco de Assis Tenório (2010) *Recent advances in partitioning clustering algorithms for interval-valued data*. EGC 2010: 19-20

- [176] De Carvalho, Francisco de A.T. De Souza Renata M. C. R. (2010): Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters* 31(5): 430-443 (2010)
- [177] De Carvalho F.A.T, Lechevallier Y, Verde R. (2006). *Symbolic clustering of large datasets*. In: Batagelj V., Bock H.H., Ferligoj A., Ziberna A (Eds). Data Science and Classification. ISBN: 3-540-34415-2. BERLIN: Springer.
- [178] De Carvalho F. De A.T., Lechevallier Y, Verde R. (2008). Clustering methods in Symbolic data analysis. In: M. Noirhomme, E. Diday. Symbolic data analysis and the SODAS software. pp. 181-204. ISBN: 9780470018835. : Wiley (USA).
- [179] de Carvalho F.A.T., Tenorio C.P.: Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems* 161(23): 2978–2999 (2010)
- [180] De Gooijer J.G. Hyndman R.J. (2006) *25 Years of Time Series Forecasting* in International Journal of Forecasting 22(3), 443-473.
- [181] De Lima, P.J.F. (1998). Non-linearities and nonstationarities in stock returns. *Journal of Business & Economics Statistics* 16, 227-236.
- [182] De A. Lima Neto E., de Carvalho Francisco De A.T.: Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis* 54(2): 333-347 (2010)
- [183] D'Esposito M.R., Palumbo F., Ragozini G. (2010). *Archetypal Symbolic Objects*. In: 45th Scientific meeting of the Italian Statistical Society Padova 16-18 Giugno 2010 Padova Società Italiana

di statistica Pag.1-8 ISBN:9788861295667 ID:3018029 Proceeding
(atti di congressi)

- [184] Di Fonso T. Lisi F. (2005) *Serie Storiche Economiche* Carocci Editore
- [185] Di Zaven A. K. Dudewicz E.J. *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods* CRC PRESS
- [186] Diebold, F.X. and J.A. Lopez (1996) *Forecast Evaluation and Combination*, in Maddala and Rao (eds.) *Handbook of Statistics*, Elsevier, Amsterdam.
- [187] Deistler M. Zinner C. (2007) *Forecasting Financial Time Series* Presentation, Canberra February 2007
- [188] Delaigle A. Hall P. (2009) *Concept of Density for Functional Data* Presentation
- [189] Delicado P. (2010) Dimensionality reduction when data are density functions *Computational Statistics Data Analysis* Volume 55, Issue 1, 1 January 2011, Pages 401-420
- [190] De Livera A.M. Hyndman R.J. Snyder R.D. (2010) Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association* (forthcoming)
- [191] Dempster M.A.H. (1974) *An Application of Quantile Arithmetic to the Distribution Problem of Stochastic Linear Programming* Bulletin of the Institute of Mathematics and its Applications

- [192] Dempster M.A.H. (1980) *Introduction to Stochastic Programming* in Stochastic Programming M. A. H. Dempster ed. Academic Press London.
- [193] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):138.
- [194] Dempster M.A.H., Papagakipapoulas A. (1980) *Computational Experience with an Approximate Method for the Distribution Problem* in Stochastic Programming M. A. H. Dempster ed. Academic Press London.
- [195] Denby L. and Mallows C. (2007) *emphVariations on the Histogram* Working Paper
- [196] Deutsch, M. & Granger, C. W. J. Terasvirta, T., (1994) The combination of forecasts using changing weights, *International Journal of Forecasting*, Elsevier, vol. 10(1), pages 47–57, June.
- [197] Dias S. Brito P. (2011) *Linear regression with histogram-valued variables* Workshop in Symbolic Data Analysis Namur, Belgium, June 2011
- [198] Diday, E. and Govaert, G. (1977) : Classification Automatique avec Distances Adaptatives. R.A.I.R.O. *Informatique Computer Science*, Vol.11, N.4, 329–349
- [199] Diday, E. (1971) Le methode des nuees dynamique, *Revue de Statistique Appliquée*, Vol.19, N.2, 19-34.
- [200] Diday E., (1980) *Optimisation en classification automatique*, INRIA.

- [201] Diday E., (1987) Une introduction à l'analyse des données symboliques, SFC, Vannes.
- [202] Diday E. Introduction l'approche symbolique en Analyse des Données. *Première Journées Symbolique-Numérique*. Université Paris IX Dauphine.
- [203] Diday E. (1993) *An Introduction to the Symbolic Data Analysis* Working Paper INRIA n.1936
- [204] Diday E. (1998) *L'Analyse des Données Symboliques: un cadre théorique et des outils*. Cahiers du CEREMADE.
- [205] Diday E. (1998) *Symbolic Data Analysis: A Mathematical Framework and Tool for Data Mining* IFCS 1998
- [206] Diday E. (2002) *An Introduction to Symbolic Data Analysis and the Sodas Software* Journal of Symbolic Data Analysis, 0 (0) ISSN 1723-5081
- [207] Diday E., (2002) An introduction to Symbolic Data Analysis and the Sodas software. *Journal of Symbolic Data Analysis*, Vol.1, N.1, International Electronic Journal
- [208] Diday E. (2006) *From Data Mining to Knowledge Mining: Symbolic Data Analysis and the Sodas Software* Computational Statistics: An advanced course on Knowledge Extraction by Interval Data Analysis, Belvedere di San Leucio (Caserta). Nova Universitas Course.
- [209] Diday E. (2008) *The State of the Art in Symbolic Data Analysis: Overview and Future*, in Symbolic Data Analysis and the SODAS software edited by E.Diday and M.Noirhomme-Fraiture, Wiley & Sons.

- [210] Diday E. (2008) Spatial classification. *DAM (Discrete Applied Mathematics)* Volume 156, Issue 8, Pages 1271-1294.
- [211] Diday E. (2010) *Symbolic Data Analysis of Complex Data: Several Directions of Research* Presentation at Compstat 2010
- [212] Diday E. (2010) *Symbolic Data Analysis Of Complex Data* CEREMADE Paris Dauphine University
- [213] Diday E. (2011) *Principal Component Analysis for Categorical Histogram Data: Some Open Directions of Research* In: Fichet B., Piccolo D., Verde R., Vichi M., "Classification and Multivariate Analysis for Complex Data Structures" Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg.
- [214] Diday E. (2011) *Symbolic data analysis for complex data* Workshop in Symbolic Data Analysis Namur 2011.
- [215] Diday E. (2011) *Principal Component Analysis for Categorical Histogram Data: Some Open Directions of Research* Classification and Multivariate Analysis for Complex Data Structures – Studies in Classification, Data Analysis, and Knowledge Organization, 2011, Part 1, 3–15,
- [216] Diday E. and Esposito F. (2003), *An introduction to Symbolic Data Analysis and the SODAS software* Intelligent Data Analysis 7(6) (2003), 583602, IOS Press.
- [217] Diday E., Lechevallier Y., Schader M., Bertrand P., Burtschy B. (1994), *New Approaches in Classification and Data Analysis*, Springer-Verlag .
- [218] Diday, E., Noirhomme,F. (2008) *Symbolic Data Analysis and the SODAS Software* Wiley-Interscience, Chichester.

- [219] Diday, E., and Simon, J.C. (1976): Clustering analysis. In *Digital Pattern Recognition*, 47-94, Springer Verlag, Heidelberg.
- [220] Diebold, F.X., Gunther, T.A. & Tay, A.S., (1997) Evaluating density forecasts. *International Economic Review*, 1574 (May), p.863-883.
- [221] Di Fonzo, Lisi (2005) *Serie Storiche Economiche* Carocci;
- [222] Di Nardo J., Tobias J.L. (2001) Nonparametric Density and Regression Estimation, *The Journal of Economic Perspectives* Vol.15 N.4 (Autumn 2001) pp.11-28
- [223] Dobb J.L. (1953) *Stochastic Processes*, John Wiley and Sons, Inc., New York, N. Y.
- [224] Domingos P. Hulten G. (2010) *A General Framework for Mining Massive Data Streams*, Working Paper
- [225] Dose, C. and Cincotti, S., 2005. Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1), p.145-151.
- [226] Downs T. Cook A. S. Rogers G. (1984) A Partitioning Approach to Yield Estimation for Large Circuits and Systems. *IEEE Transactions on Circuits and Systems* 31, 1984
- [227] Drago C. (2009) *Symbolic Data Analysis of Complex Data* Presentation Department of Mathematics and Statistics, University of Naples "Federico II"
- [228] Drago C. (2010) *Forecasting Interval Time Series Using Combinations and Hybrid Models* Institute de Investigacion Tecnologica

- (IIT), Escuela Tecnica Superior de Ingenieria, ICAI, Universidad Pontificia Comillas, Madrid (Spain) 16 september 2010
- [229] Drago C. (2011) *Beanplot Time Series Analysis: Visualization, Clustering and Forecasting* Presentation Department of Mathematics and Statistics, University of Naples "Federico II"
- [230] Drago C., Lauro N.C., Marchitelli M. Scepi G. (2009) "Ex Ante Evaluation of Public Services: A Multivariate Statistical Interval Data based Approach" Working Paper
- [231] Drago C. Irace D. (2004) *Analisi dei dati intervallari: l'approccio dei dati simbolici* Working Paper
- [232] Drago C. Irace D. (2005) *Analisi delle Serie Storiche finanziarie: comparazione tra metodi previsivi* Working Paper
- [233] Drago C., Lauro C.N., Scepi G. (2009) *Visualization and Functional Analysis of Time Varying Histogram Data* Wienerwaldhof, Vienna (Austria), Workshop in Symbolic Data Analysis, October 2009
- [234] Drago C., Lauro C., Scepi G. (2010) *Visualizing and Forecasting Beanplot Time Series* Working Paper
- [235] Drago C., Lauro C., Scepi G. (2011) *Beanplot Data Analysis in a Temporal Framework* Working Paper
- [236] Drago C., Scepi G. (2010) *Forecasting by Beanplot Time Series* Electronic Proceedings of Compstat/, Springer Verlag, p.959-967, ISBN 978-3-7908-2603-6
- [237] Drago C., Scepi G. (2010) *Visualizing and exploring high frequency financial data: beanplot time series* accettato su : New Perspectives in Statistical Modeling and Data Analysis, Springer

Series: Studies in Classification, Data Analysis, and Knowledge Organization, Ingrassia, Salvatore; Rocci, Roberto; Vichi, Maurizio (Eds), ISBN: 978-3-642-11362, atteso per novembre 2010.

- [238] Drucker P.F (1992) *Managing for the Future: The 1990s and Beyond* Dutton Adult, 1992.
- [239] Du J. (2002) *Combined Algorithms for Constrained Estimation of Finite Mixture Distributions with Grouped Data and Conditional Data* Master of Science dissertation in Statistics McMaster University Hamilton Ontario
- [240] Dubois D., Prade, H.(1988) *Possibility Theory An Approach to Computerised Processing of Uncertainty* Plenum Press, New York.
- [241] Dunis C., Gavridis M., Harris A., Leong S., Nacaskul P. (1998) An Application of genetic algorithms to high frequency trading models: a case study, in: *Nonlinear Modelling of High Frequency Financial Time Series*, Dunis C. Zhou B. (Eds.), Wiley, 247-278
- [242] Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4, 95104.
- [243] Durbin, J. (2004) *Introduction to state space time series analysis*. In personal unpubl. Cambridge University Press, p. 325. Available at: <http://eprints.ucl.ac.uk/18379/>.
- [244] Durbin J., and Koopman S.J. (2001) *Time Series Analysis by State Space Methods* Oxford University Press
- [245] Ecker K. Ratschek H. (1972) *Intervallarithmetik mit Wahrscheinlichkeitsstruktur* Angewandte Informatik Elektron Datenverarbeitung) 14 (1972)

- [246] Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S., (2002) *Background and foreground modeling using nonparametric kernel density estimation for visual surveillance* Proceedings of the Jul. 2002, Volume 90, Issue 7, 1151 - 1163
- [247] Elliott, G. & Timmermann, A., (2008). Economic Forecasting. *Journal of Economic Literature* 46(1), p.3-56.
- [248] Enders, W. (1995), *Applied econometric time series*, John Wiley & Sons.,Inc., New York
- [249] Engle R. F. (1982) *Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation* *Econometrica*, 50(4),9871008.
- [250] Engle, R. F. (2000), *The Econometrics of Ultra-high-frequency Data* *Econometrica*, 68: 122. doi: 10.1111/1468-0262.00091
- [251] Engle, R. F., Granger, C. W. J. (1987) "Co-integration and error correction: Representation, estimation and testing", *Econometrica*, 55(2), 251-276.
- [252] Engle R.F., Manganelli S. (2004) *Journal of Business and Economic Statistics*. October 1, 2004, 22(4): 367-381.
- [253] Engle R.F, Russell J.R. (1998) *Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data*. *Econometrica*, 1998, 66(5), pp. 1127-62.
- [254] Engle R.F., Russell, J.R. *Analysis of High Frequency Financial Data* Working Paper
- [255] Engle R.F., Russell, J.R. (2009) *Analysis of high frequency data* In: Ait Sahalia,Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics Vol.1 Tools and Techniques*

- [256] Epanechnikov, V.A. (1969) Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications* 14: 153158.
- [257] Esty, W.W. & Banfield, J.D. 2003. The box-percentile plot. *Journal of Statistical Software*, 8, 114.
- [258] Fabbri G. Orsini R. (1993) *Reti Neurali per le Scienze Economiche* Franco Muzzio Editore.
- [259] Fair R. (2004) *Estimating How the Macroeconomy Works*, Harvard University Press.
- [260] Falkenberry T.N. (2002) High Frequency Data Filtering. Tick Data White Paper
- [261] Fantazzini D. Rossi E. (2005) *Asymmetric Periodic Models for High Frequency Data Analysis* Presented (Poster Session) at the Conference on “Changing Structures in International and Financial Markets and the Effects on Financial Decision Making”, Venice, Italy, June 2-3.
- [262] Faraway J. J. and M. Jhun (1990), Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association* 85, 11191122.
- [263] Fadallah A.(2011) *Highest Density Regions for Univariate and Bivariate Densities* Econometrics and Operations Research. Thesis: Bachelor (FEB 23100) Rotterdam, 6 July 2011
- [264] Farmer J.D. Gillemot L. Lillo F. Mike S. and Sen A. (2004) *What really causes large price changes?* *Quantitative Finance*, 4 (2004), pp. 383397.

- [265] Favero A. (2001) *Applied Macroeconometrics* Oxford University Press pages 51-238
- [266] Fawcett, N. W. P. and Hendry, D. F. (2007) *Learning and Forecasting: UK M1 revisited*, mimeo, University of Oxford, Oxford.
- [267] Fernandes M., de Sa Mota B. and Rocha G. (2005), A multivariate conditional autoregressive range model, *Economics Letters*, 86, issue 3, p. 435-440.
- [268] Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995) *Knowledge extraction using stochastic matrices: Application to elaborate a fishing strategy* Proc. Ordinal and Symbolic Data Analysis. Paris ; Diday, Lechevallier, Opitz edit. Springer Studies in Classification.
- [269] Ferson, S. (2002) *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers* Lewis Publishers, Boca Raton, Florida.
- [270] Figueiredo L.H., Stolfi J.: *Affine arithmetic* <http://www.ic.unicamp.br/~stolfi/EXPORT/projects/affine-arith/>.
- [271] Fischer A. (2011) Predictability of Asset Returns - Lecture Notes
- [272] Fox J. (2002) *Nonlinear Regression and Nonlinear Least Squares* Working Paper
- [273] Fowlkes, E. B., and Mallows C. L. (1983). A Method for Comparing Two Hierarchical Clusterings *J. Am. Stat. Assoc.* 78:553-569.
- [274] Fraley C. (1996) *Algorithms for Model Based Gaussian Hierarchical Clustering* Technical Report No. 311 Department of Statistics University of Washington

- [275] Fraley C, Raftery AE (1998). *How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis* Computer Journal, 41, 578-588.
- [276] Fraley C, Raftery AE (1999). *mclust: Software for Model-based Cluster Analysis*. Journal of Classification, 16, 297-306.
- [277] Fraley, C., Raftery A.E. (2002) *Model-based Clustering, Discriminant Analysis and Density Estimation* Journal of the American Statistical Association, 97, 611-631.
- [278] Fraley C., Raftery A.E. (2006) *Mclust Version 3 for R: Normal Mixture Modeling and Model-Based Clustering* September 2006 Working paper.
- [279] Fraley, C., Raftery A.E. (2007) *Model-based methods of Classification: Using the mclust Software in Chemometrics* Journal of Statistical Software volume 18, issue 6.
- [280] Francois R. 2011 *Kernel density estimator: Illustration of the kernels* R Graphics <http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=30>
- [281] Fricks J. (2007) *Time Series II Frequency Domain Methods* Astrostatistics Summer School Penn State University University Park, PA 16802 June 8, 2007
- [282] Friedman J.H, Hastie T., Tibshirani R. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Second Edition, Springer-Verlag, Heidelberg, 2009
- [283] Fritz H., García Escudero L. Mayo-Isacar A. (2011) *tclust: An R Package for a Trimming Approach to Cluster Analysis* Working paper.

- [284] Fryer, M. J. (1977): *A review of some non-parametric methods of density estimation*, Journal of the Institute of Mathematics Applications, 20, 335-354.
- [285] Gaber, M.M., Zaslavsky, A. and Krishnaswamy, S. (2005). Mining data streams: a review. *Techniques*, 34(2), p.18-26. Available at: <http://portal.acm.org/citation.cfm?id=1083789>.
- [286] Galli F. (2003) *Econometria dei dati finanziari ad alta frequenza* Ph.D Thesis in Economics, University of Pavia.
- [287] Gantz J.F., Reinsel D., Chute C., Schlichting W., Minton S., Toncheva A., and Manfrediz A. (2008) *The expanding digital universe: An updated forecast of worldwide information growth through 2011*. The Diverse and Exploding Digital Universe. IDC White Paper. Bohn, Germany: IDC. Retrieved from <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>
- [288] Gantz J.F and Reinsel D. (2009) *As the economy contracts, the digital universe expands* IDC white paper sponsored by EMC May 2010
- [289] Gantz J.F. Reinsel D.(2010) *The digital universe decade are you ready?* IDC iView, sponsored by EMC May 2010.
- [290] Gao, J. and Cao, Y. and Tung, W. and Hu, J. (2007) *Multiscale analysis of complex time series* Wiley
- [291] García-Ascanio, C. Maté C. 2010 "Electric power demand forecasting using interval time series: A comparison between VAR and iMLP", *Energy Policy*. vol. 38, no. 2, pp. 715-725, February 2010.

- [292] García-Escudero, L. A., Gordaliza, A., and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434–449
- [293] Gelman A. (2009) *What's wrong with a kernel density?* Blog: Statistical Modeling, Causal Inference, and Social Science, November 25 2009
- [294] Gelman, A., and Stern, H. (2006), The Difference Between Significant and Not Significant is not Itself Statistically Significant, *The American Statistician*, 60, 328331.
- [295] Gershenfeld N.A. (1999) *The nature of mathematical modeling* Cambridge Univ Press
- [296] Gettler Summa M. Goldfarb B. *About some crucial issues in temporal data analysis* Working Paper, Communications Cere-made.
- [297] Gettler-Summa, M., Pardoux, C. (2000) in Noirhomme-Fraiture, Rouard M. (2000) "Symbolic Approaches for Three-way Data" (Chapter 12), In Bock, H. H. Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 342-354
- [298] Gettler Summa M. et al. (2006) Multiple Time Series: New Approaches and New Tools in Data Mining Applications to Cancer Epidemiology. *Revue Modulad* n.34
- [299] Gettler-Summa M. Frédérick V. (2010) *Editing and Processing Complex Data* 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)

- [300] Gherghi M. Lauro C. (2002) *ppunti di analisi dei dati multidimensionali : Metodologie ed esempi* Università di Napoli Federico II. Dipartimento di Matematica e Statistica, Edisu Napoli
- [301] Ghysels E. (2005) *MIDAS Estimation: Applications in Finance and Macroeconomics* 16th (EC)² Conference Istanbul 2005
- [302] Ghysels E. Santa Clara P. Valkanov R. (2004) *The MIDAS Touch: Mixed Data Sampling Regression Models* Working Paper
- [303] Ghysels E., Sinko A., Valkanov R. (2007) MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26 (1), 5390
- [304] Ghosh, A. and Bera, A. K. 2001 *Neyman's Smooth Test and Its Applications in Econometrics*. Available at SSRN: <http://ssrn.com/abstract=272888> or doi:10.2139/ssrn.272888
- [305] Gibbs, A.L. and SU, F.E. (2002): On choosing and bounding probability metrics. *Intl. Stat. Rev.* 7 (3), 419–435.
- [306] Gilbert P.D. and Meijer E. (2006). *Money and Credit Factors* Working Papers 06-3, Bank of Canada.
- [307] Gioia F.(2001) *Statistical Methods for Interval Variables*, Ph.D. thesis, Dep. of Mathematics and Statistics, University Federico II Naples, in Italian.
- [308] Gioia F. (2006) *Basic Statistics and Interval Algebra*. Presented in the Advanced Course on Knowledge Extraction by Interval Data Analysis, at Belvedere di San Leucio, Caserta, 27-29 novembre 2006
- [309] Gioia F.(2008) *Portfolio Selection Models with Interval Data* Decision in Economics and Finance Manuscript

- [310] Gioia F. Lauro C. (2005) *Basic statistical methods for interval data* Statistica Applicata, 17, 1,75–104.
- [311] Giordano G., Palumbo F. (1999), *A New Statistical Quality Control Tool Based on PCA of Interval Data*, in Book of short papers, Cladag '99, Roma, pp 197–200.
- [312] Giudici P. (2006) *Data Mining - Metodi informatici, statistici e applicazioni* 2/ed McGraw Hill
- [313] Giudici P., Figini S. (2009) *Applied Data Mining for Business and Industry* Wiley, London
- [314] Giudici P., Heckerman D., Whittaker J. (2001) Statistical models for data mining, in "Knowledge discovery and data mining", 5.
- [315] González-Rivera G. Carlos Maté (2007) *Forecasting histogram-valued time series (HTS) in financial markets. Applications to stock indices* Department of Economics, Econometrics Colloquia Nov.26-2007, University of California, Riverside
- [316] Goodfriend M. (1992) Information Aggregation Bias *The American Economic Review* Vol. 82, No. 3, Jun., 1992
- [317] Goupil F., Touati M., Diday E., Van Der Veen H. (2000) Working Paper
- [318] Gonzáles Rivera G. Arroyo J. (2011) *Autocorrelation function of the daily histogram time series of SP500 intradaily returns* Working Paper
- [319] Gonzáles Rivera G. Arroyo J. (2011) Time Series Modeling of Histogram-valued Data The Daily Histogram Time Series of

- SP500 Intradaily Returns. *International Journal of Forecasting*
Article in Press, doi:10.1016/j.ijforecast.2011.02.007
- [320] Gordon A.D. (1999) A survey of constrained classification *Computational Statistics & Data Analysis* Volume 21, Issue 1, January 1996, Pages 17-29
- [321] Granger C. (1981) "Some Properties of Time Series Data and Their Use in Econometric Model Specification", *Journal of Econometrics* 16: 121-130.
- [322] Granger C.W.J., Ramanathan R. (1984) Improved methods of combining forecasts, *Journal of Forecasting*,3,2,197–204, Wiley Online Library
- [323] Granville V. (2011) *Data Science by Analyticbridge* E–Book, first draft October 2011
- [324] Grassia M.G. Lauro C.N. Scepi G. (2004) *L'Analisi dei dati ad intervallo nell'ambito della Qualità* in Davino C. Lauro N.C. (a cura di) (2004) *Data Mining e Analisi Simbolica*. Franco Angeli editore
- [325] Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall.
- [326] Griethe H. and Schumann H. (2006) *The Visualization of Uncertain Data: Methods and Problems* In Proceedings of SimVis '06, 2006.
- [327] Grubbs F. (1969), Sample Criteria for Testing Outlying Observations, *Annals of Mathematical Statistics* pp. 27-58.

- [328] Grubbs, F. (1969), Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 11(1), pp. 1-21. February
- [329] Gunopulos D. (2011) *Dimensionality Reduction Techniques* Lecture-Notes
- [330] Gupta A. Santini S. (2000) Toward feature algebras in visual databases: The case for a histogram algebra. *Proceedings of the IFIP Working Conference on Visual Databases (VDB5)*, Fukuoka (Japan), Citeseer.
- [331] Guo, J., Li, W., Li, C., and Gao, S. (2011). Standardization of interval symbolic data based on the empirical descriptive statistics. *Computational Statistics & Data Analysis*.
- [332] Hallin, M. & Puri, M.L. (1992). *Rank Tests for Time Series Analysis*, A Survey Papers 9210, Universite Libre de Bruxelles - C.E.M.E.
- [333] Hamilton J.D (1994) *Time Series Analysis* Princeton University Press
- [334] Hamilton J.D. (2005) *Regime Switching Models* Palgrave Dictionary of Economics
- [335] Hammer Ø(2011) *PAST PAleontological STatistics Version 2.12, Reference manual* Natural History Museum University of Oslo
- [336] Han, A, Hong, Y. & Wang, S. (2009). *Autoregressive conditional models for interval-valued time series data* Working paper.
- [337] Han, A. and Hong, Y. and Lai, KK and Wang, S. (2008) Interval time series analysis with an application to the Sterling-Dollar exchange rate *Journal of Systems Science and Complexity*, 21, 4, 558-573, Springer

- [338] Hansen B.U (2009) *Lecture Notes on Nonparametrics* University of Wisconsin Spring 2009
- [339] Hardin, J. and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44(4), 625–638.
- [340] Harrell F., Banfield J., Hyndman R.J., Adler D., (2011) *The Boxplot Friends* R Graphics <http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=102>
- [341] Harris H.H. (2011) *What is "Data Science" Anyway?* Presentation September 26 2011
- [342] Hart J.D., Vieu P. *Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data* Ann. Statist. Volume 18, Number 2 (1990), 873-890.
- [343] Harvey A.C. (1990) *The Econometric Analysis of Time Series* MIT Press
- [344] Harvey A.C. (2010) *Dynamic Distributions and Changing Copulas* Forthcoming in the Journal of Empirical Finance
- [345] Harvey A. Oryshchenko V. (2010) *Kernel density estimation for time series data* Working Paper
- [346] Hastie, T. J. (1992) Generalized additive models. Chapter 7 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth Brooks/Cole.
- [347] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall.

- [348] Hautsch N. (2007) *Point Process Models for Financial High-Frequency Data* CCFEA Summer School 2007, University of Essex, September 6, 2007
- [349] Hautsch N. Kyj L.M. Malec P. (2011) *The Merit of High-Frequency Data in Portfolio Allocation*. Working Paper Goethe Universität Frankfurt Am Main
- [350] He K., Meedn G. (1997) Selecting the number of bins in a histogram: A decision theoretic approach. *Journal of Statistical Planning and Inference* 61(1): 49-59
- [351] He Y., Hong Y., Ai Han A., Wang S. (2011) Forecasting of Interval-valued Crude Oil Prices with Autoregressive Conditional Interval Models. Working Paper.
- [352] He, L.T. and C. Hu (2009). Impacts of Interval Computing on Stock Market Variability Forecasting. *Computational Economics* 33, 263-276.
- [353] He, A.W.W., Kwok, J.T.K, and Wan, A.T.K. (2010). An empirical model of daily highs and lows of West Texas Intermediate crude oil prices. *Energy Economics* 32, 1499–1506.
- [354] Hendry, D. F. (2006) Robustifying Forecasts from Equilibrium-Correction Models, *Journal of Econometrics*, 135 (1-2), 399-426
- [355] Hendry, D. F. and Hubrich, K. (2006) Forecasting aggregates by disaggregates. European Central Bank Working Paper 589.
- [356] Hendry, D. F. and Hubrich, K. (2010) Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business and Economic Statistics*

- [357] Hennig C. (2009) *Dissolution and isolation robustness of fixed point clusters* in Okada, A.; Imaizumi, T.; Bock, H.-H.; Gaul, W. (Eds.): *Cooperation in Classification and Data Analysis Proceedings of Two German-Japanese Workshops*. Springer, Berlin 2009.
- [358] Henry M. Zaffaroni P. 2003 *The Long Range Dependence Paradigm for Macroeconomics and Finance* 419–438, Birkhauser
- [359] Hey T. Tansley S. Tolle K. (2010) *The Fourth Paradigm: Data-Intensive Scientific Discovery* Microsoft Research
- [360] Hickey, T., Ju, Q., and Van Emden, M. H. (2001) Interval arithmetic: From principles to implementation. *Journal of the ACM* 48(5): 1038–1068.
- [361] Hilbert M., Lopez P. *The World's Technological Capacity to Store, Communicate, and Compute Information* Science Express. Published online 10 February 2011.
- [362] Hintze, J. L., and Nelson. R.D. (1998) *Violin Plots: A Box Plot-Density Trace Synergism* The American Statistician 52(2):181-84
- [363] Holton, G. (2003). *Value-at-Risk: Theory and Practice*. Academic Press. ISBN 978-0123540102.
- [364] Horenko I. (2010) *On Clustering of Non-stationary Meteorological Time Series* Dyn. of Atm. and Oc. , 49, pp.164-187
- [365] Horenko I. (2010) *Finite Element Approach to Clustering of Multidimensional Time Series* SIAM J. Sci. Comp. 32 (1), pp. 62-83
- [366] Hornik K. (2005) A CLUE for CLUster ensembles. *Journal of Statistical Software*, 14, 12, 1–25, 2005, Citeseer.

- [367] Howell, D. C. (2007) The analysis of missing data In Outhwaite, W. and Turner, S. *Handbook of Social Science Methodology* London: Sage.
- [368] Hsiao, C. and Fujiki, H. (1998) *Nonstationary Time-Series Modeling versus Structural Equation Modeling: With an Application to Japanese Money Demand* Monetary and Economic Studies, Institute for Monetary and Economic Studies, Bank of Japan, vol. 16(1), pages 57–79, May.
- [369] Hsu, H.L. and Wu B. (2008). Evaluating Forecasting Performance for Interval Data. *Computers and Mathematics with Application*, 56, 2155–2163.
- [370] Hu, C., Baker Kearfott, R., Korvin, A. de, Kreinovich, V. (2008) *Knowledge Processing with Interval and Soft Computing* Springer
- [371] Hu, C. and He L.T. (2007) An application of interval methods to stock market forecasting. *Reliable Computing*, 13(5), 423–434.
- [372] Huang T.M, Kecman V., Kopriva I. (2006) *Kernel Based Algorithms for Mining Huge Data Sets* Series Studies in Computational Intelligence, Vol. 17 Springer Verlag, Berlin, Heidelberg
- [373] Huber P.J. 1981 *Robust Statistics* John Wiley & Sons, New York, 1981.
- [374] Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers* 116, 770–799, 800–808.
- [375] Hutchinson J. *Big data to get even bigger in 2011* Computer-world Australia

- [376] Hickey, T., Ju, Q., Van Emden, M. H. (2001) Interval arithmetic: From principles to implementation. *Journal of the ACM*, 48, 5, 1038–1068.
- [377] Hyndman R.J (1995) *The Problem with Sturges rule for constructing histograms* Working paper
- [378] Hyndman R.J. (1996) Computing and graphing highest density regions. *American Statistician*, 120–126, JSTOR
- [379] Hyndman R.J. (2006) Another look at forecast-accuracy metrics for intermittent demand. *Foresight* June 2006 Issue 4
- [380] Hyndman, R.J., Akram, Md., and Archibald, B. (2008) The admissible parameter space for exponential smoothing models. *Annals of Statistical Mathematics*, 60(2), 407426.
- [381] Hyndman R.J., and Billah B. (2003) Unmasking the Theta method *International Journal of Forecasting*, 19, 287–290
- [382] Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26(3). (2008).
- [383] Hyndman R.J, King R.J., Pitrun I. and Billah B. (2005) *Local linear forecasts using cubic smoothing splines* Australian and New Zealand Journal of Statistics, 47(1), 87-99
- [384] Hyndman, R.J., Koehler, A.B., Snyder, R.D., and Grose, S. (2002) A state space framework for automatic forecasting using exponential smoothing methods *International Journal of Forecasting* 18(3), 439454.

- [385] Hyndman, R.J., Koehler, A.B., Ord, J.K., and Snyder, R.D. (2008) *Forecasting with exponential smoothing: the state space approach* Springer-Verlag.
- [386] Hyndman, R, and A Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22, no. 4: 679-688.
- [387] Hyndman, R. J. and Fan, S. (2008), *Density forecasting for long-term peak electricity demand* 29 Working paper 6/08, Department of Econometrics & Business Statistics, Monash University. <http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2008/wp6-08.pdf>
- [388] Hyndman, King, Pitrun and Billah (2005) Local linear forecasts using cubic smoothing splines. *Australian and New Zealand Journal of Statistics*, 47(1), 87–99.
- [389] Hu Y., Chou R.J. (2003) A Dynamic Factor Model. *Journal of Time Series Analysis*
- [390] Imakor M., Billard L., Diday E. (2006) *PLS Symbolic Approach* Working Paper
- [391] Ingrassia S., Rocci R. (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis* 51, 5339–5351.
- [392] Ingrassia S., Greselin F., Morlini I. (2008) *Modelli di Mistura e Algoritmo EM* Course on Computational Statistics, Nova Universitas, Macerata 25–26 September 2008
- [393] Irpino, A. (2006) "Spaghetti" PCA analysis: An extension of principal components analysis to time dependent interval data. *Pattern Recognition Letters* 27(5), p.504-513.

- [394] Irpino, A. (2009) "Tecniche di raggruppamento per dati ad intervallo" Presentation.
- [395] Irpino, A. and Romano, E. (2007): *Optimal histogram representation of large data sets: Fisher vs piecewise linear approximations*. RNTI E-9, 99–110.
- [396] Irpino A., Tontodonato V. (2006). Clustering reduced interval data using Hausdorff distance. *Computational Statistics*. vol. 21, pp. 271-288 ISSN: 0943-4062.
- [397] Irpino A, Verde R. (2007). *Dynamic clustering of histogram data: using the right metric*. In: Brito P., Bertrand P., Cucumel G., De Carvalho F. (Eds). Selected contributions in Data Analysis and Classification. vol. XIII, pp. 123-134. ISBN: 978-3-540-73558-8. Heidelberg: Springer.
- [398] Irpino A. Verde R. (2008), *Comparing Histogram data using a MahalanobisWasserstein distance* In: Brito, P. (eds.) COMP-STAT 2008. PhysicaVerlag, Springer, Berlin, 7789.
- [399] Irpino A, Verde R., Lechevallier Y. (2006). *Dynamic clustering of histograms using Wasserstein metric*. In: Rizzi A., Vichi M. (Eds). Advances in Computational Statistics. pp. 869-876. ISBN 978-3-7908-1708-9. Heidelberg: Physica-Verlag
- [400] Jaccard P. (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37: 547579
- [401] Jackson, C. H. (2008) Displaying uncertainty with shading. *The American Statistician*, 62(4):340-347. <http://www.mrcbsu.cam.ac.uk/personal/chris/papers/denstrip.pdf>

- [402] Jain A.K. (2010) Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*,31,8, 651–666, 2010,Elsevier.
- [403] Jain, A.K. and Murty, M.N. and Flynn, P.J.(1999) Data clustering: a review, *ACM computing surveys (CSUR)*,31,3,264–323,1999,ACM
- [404] Johansen,S. (1988), "Statistical Analysis of Cointegrating Vectors", *Journal of Economic Dynamics and Control* 12, 231–254
- [405] Johansen ,S. (1991), "Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregressive Models", *Econometrica* 59, 1551–1580
- [406] Johansen, S. (1994), "The Role of the Constant and Linear Terms in Cointegration Analysis of Nonstationary Variables", *Econometric Reviews* 13(2)
- [407] Johansen, S. and Juselius K. (1990), "Maximum Likelihood Estimation and Inference on Cointegration, with Applications to the Demand for Money", *Oxford Bulletin of Economics and Statistics* 52, 169–210
- [408] Johnson C.R. (2004) *Top Scientific Visualization Research Problems* In IEEE Computer Graphics and Applications, vol. 24,no. 4, pp. 13–17
- [409] Johnson N.F., Jefferies P., Ming Hui P. (2000) *Financial market complexity* Oxford University Press
- [410] Jolliffe I. (2005) *Statistical Models for Probabilistic Forecasting* AMS Short Course on Probabilistic Forecasting, 9th January 2005

- [411] Jones, M.C. (1992) Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44, 721-727.
- [412] Jones M. C. (1993). Kernel density estimation when the bandwidth is large. *Australian Journal of Statistics* 35(3), pp. 319-326
- [413] Jones, M.C., Marron, J.S., and Sheather, S.J. (1996) *A brief survey of bandwidth selection for density estimation* Journal of the American Statistical Association, 91, 401-407
- [414] Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk (3rd ed.)*. McGraw-Hill
- [415] Kaminska I. (2008) *High frequency traders do risk better* ft.com alphaville September 8 2008.
- [416] Kampstra P.(2008) *Beanplot: A boxplot alternative for visual comparison of distributions*, Journal of Statistical Software Code Snippets, 28(1), 2008.
- [417] Kang H. (1986) Unstable Weights in the Combination of Forecasts *Management Science* June 1986 vol. 32 no. 6 683–695
- [418] Karr, A. (1991), *Point Processes and Their Statistical Inference* second ed. Dekker, NY.
- [419] Katkovnik V., Shmulevich I. (2000) *Pattern Recognition Letters* 23 (2002) 1641-1648
- [420] Keim D. B. (1983) Size-Related Anomalies and Stock Return Seasonality: Further Empirical Evidence, *Journal of Financial Economics* 12 (1983)

- [421] Keogh E., Chu S., Hart D., Pazzani M. (2004) Segmenting time series: a survey and novel approach, in: M. Last, A. Kandel, H. Bunke (Eds.), *Data Mining in Time Series Databases*, World Scientific, Singapore, 2004.
- [422] Lin, J., Keogh, E., Li, W. & Lonardi, S. (2007). Experiencing SAX: A Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery Journal*. p. 107-144. (This work also appears in: Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003). A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13.)
- [423] Kearfott V. Kreinovich R. B. (1996) *Applications Of Interval Computations*, Kluwer Academic Publishers.
- [424] Kim, J. and Scott, C. *Robust kernel density estimation* IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), April 2008.
- [425] Kisimbay T (2010) The use of encompassing tests for forecast combinations *Journal of Forecasting*, John Wiley & Sons, Ltd., vol. 29(8), pages 715-727, December.
- [426] Kneip A., Utikal K.J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* 96, 519-532.
- [427] Knoth, S. and W. Schmid, (2004) Control charts for time series: a review. in *Frontiers of Statistical Quality Control*, Lenz, H.-J. and Wilrich, P.-Th. (Eds.) 7.
- [428] Knott G.D. (2000), *Interpolating cubic splines*. Springer. p. 151

- [429] Koenker R. (1996) *Reproducible Econometric Research*, Department of Econometrics, University of Illinois, Urbana-Champaign, IL, Tech. Rep.
- [430] Koenker R. and Basset G. (1978) *Regression quantiles*, *Econometrica*, 46, 33-50.
- [431] Kovalerchuk, B. Vityaev E. 2000 *Data Mining in Finance* Spring
- [432] Krivoruchenko, M.I. et al., (2004). Modeling stylized facts for financial time series. *Physica A: Statistical Mechanics and its Applications*, 344(1-2), p.32.
- [433] Kubica B.J. and Malinowski K. (2006) Interval Random Variables and Their Application in Queueing Systems with Long-Tailed Service Times. Soft Methods for Integrated Uncertainty Modeling *Advances in Soft Computing*, 2006, Volume 37/2006, 393-403
- [434] Kulpa Z. (2001) Diagrammatic Representation for Interval Arithmetic *Linear Algebra and its Applications* (423) 55-80
- [435] Kulpa, Z. (2004) Designing diagrammatic notation for interval analysis. *Information Design Journal* vol. 12, n° 1, 2004, p. 52-62.
- [436] Kulpa Z. (2006) *Diagrammatic Interval Analysis with Applications*. IPPT PAN Reports 1/2006, xvi+232 pp., Warsaw 2006.
- [437] Kunitomo, N. (1992). Improving the Parkinson Method of Estimating Security Price Volatilities. *Journal of Business* 65, 295-302.
- [438] Kusnetzky, D. (2010) *What is "Big Data?"*. ZDNet. <http://blogs.zdnet.com/virtualization/?p=1708>

- [439] Kunst R. (2007) *Seasonality in high frequency data – Austrian electricity spot price* Working Paper
- [440] Kyle, A. S., Obizhaeva, A. A. and Tuzun, T. *Trading Game Invariance in the TAQ Dataset* (March 8, 2010). Available at SSRN: <http://ssrn.com/abstract=1107875>
- [441] Laxman, Srivatsan, and P S Sastry. 2006. A survey of temporal data mining. *Sadhana* 31, no. 2: 173-198.
- [442] Lauro N.C. (1996), *Computational statistics or statistical computing, is that the question?*, *Computational Statistics & Data Analysis* 23: 191-193.
- [443] Lauro N.C. (2004) *Factorial Conjoint Analysis: Classification and Related Issues* Presentation, 2004.
- [444] Lauro N.C. and Balbi S. (2003) *Time Dependent Non Symmetrical Correspondence Analysis* Working Paper
- [445] Lauro N.C. Palumbo F. (2000) "Principal component analysis of interval data: a symbolic data analysis approach". *Computational Statistics* v15 i1. 73-87
- [446] Lauro N.C. Palumbo F. (2003) *Some results and new perspectives in Principal Component Analysis for interval data* in Atti del Convegno CLADAG'03 Gruppo di Classificazione della Società Italiana di Statistica. Bologna 24-26 Settembre 2003, pp. 237-244, relazione invitata per sessione plenaria.
- [447] Lauro N.C. Palumbo F. (2005) Principal component analysis for non precise data, *New Developments in Classification and Data Analysis*. In: Vichi M., Monari P., Mignani S., Montanari A. Editors. Series: Studies in Classification, Data Analysis, and Knowledge Organization. (vol. X, pp. 173-184). ISBN: 3-540-23809-3.

- [448] Lauro N.C., Palumbo F. Iodice D'Enza A. (2004) Visualizzazione ed ordinamento di oggetti simbolici. in Davino C. Lauro N.C. (a cura di) *Data Mining e Analisi Simbolica*. Franco Angeli editore
- [449] Lauro N.C. and Verde R. (2009) *Symbolic Data Analysis: A New Tool in Data Mining* Working Paper
- [450] Lauro N.C., Verde R., Irpino A. (2008). *Principal components analysis of symbolic data described by intervals*. In: Noirhomme M., Diday E. *Symbolic data analysis and the SODAS software*. pp. 279-312: Wiley (USA).
- [451] Lawrence, R., (1997). *Using Neural Networks to Forecast Stock Market Prices*. Methods, Working Paper
- [452] Lebart L., Morineau A., Piron M. (1995) *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris
- [453] Lequeux P. (eds.) (1999) *Financial Markets Tick by Tick* Wiley & Sons
- [454] Lev Ram M. (2011) *Why big data is suddenly sexy* CNN Money July 15 2011 <http://tech.fortune.cnn.com/2011/07/15/why-big-data-is-suddenly-sexy/>
- [455] Liao, T. W. (2005). Clustering of time series dataa survey. *Pattern Recognition* 38 (11), 1857–1874.
- [456] Lillo F. (2010). *An Introduction to High Frequency Finance and Market Microstructure* Presentation Pisa, February 11, 2010
- [457] Lim, K.P., Brooks, R.D. and Hinich, M.J. (2008) Nonlinear serial dependence and the weak-form efficiency of Asian emerging stock

- markets, *Journal of International Financial Markets, Institutions and Money* 18(5), 527–544.
- [458] Lima Neto, E. A. and F. d. A. T. De Carvalho (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis* 52, 15001515.
- [459] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. (2003). *A Symbolic Representation of Time Series, with Implications for Streaming Algorithms*. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13.
- [460] Ling T. He and Chenyi Hu (2009) Impacts of Interval Computing on Stock Market Variability Forecasting. *Computational Economics* 33, 3 (April 2009), 263-276.
- [461] Linoff S. Berry J.A. (2011) *Data Mining Techniques for Marketing, Sales and Customer Relationship Management* (Third Edition) John Wiley & Sons
- [462] Lippi M. Thornton D.L. (2004) *A Dynamic Factor Analysis of the Response of U.S. Interest Rates to News* Economics Working Papers Federal Reserve of St.Louis
- [463] Lipkovich I., Smith E.P. (2010) *Model Based Cluster and Outlier Analysis* Working Paper
- [464] Little, R.J.A., Rubin, D.B. (1987) *Statistical analysis with missing data*. New York, Wiley
- [465] Li Q., Racine J.S. (2007) *Nonparametric Econometrics Theory and Practice*, Princeton University Press

- [466] Lo A.W. MacKinlay A.C. (1999) *A Non- Random Walk Down Wall Street* Princeton: Princeton University Press.
- [467] Lo, A. W, Mamaysky H., and Wang. Y. (2000) *Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation*. Social Science Research Network. SSRN. <http://ssrn.com/paper=228099>.
- [468] Loukides M. (2010) *What is the Data Science?* O'Reilly Radar Insight Analysis and Research on Emerging Technologies, 2 June 2010 <http://radar.oreilly.com/2010/06/what-is-data-science.html>
- [469] Lütkepohl H. (2005) *New Introduction to Multiple Time Series Analysis* Springer.
- [470] Lütkepohl, H. and Krätzig, M. (2004) *Applied Time Series Econometrics*, Cambridge University Press, Cambridge, 2004
- [471] Lütkepohl H. & Xu F. (2009) *The Role of the Log Transformation in Forecasting Economic Variables* CESifo Working Paper Series 2591, CESifo Group Munich.
- [472] Lyman P., Varian H. (2003) *How Much Information?* School of Information, Management and Systems, University of California at Berkeley 2003
- [473] Luo A., Kao D., and Pang A. (2003), Visualizing spatial distribution data sets. *Proceedings of the Symposium on Data visualization* pages 2938, 2003.
- [474] Madsen M.(2011) *The Mythology of Big Data* O'Reilly Strata Conference February 1-3 2011 Santa Clara US
- [475] Mahmoud, E., (1984). Accuracy in forecasting: A survey. *Journal of Forecasting*, 3(2), p.139-159.

- [476] Maia, A.L.S., De Carvalho,F., Ludermer,T.B., (2008) Forecasting models for interval-valued time series. *Neurocomputing* 71(1618), 33443352.
- [477] Maia A.L.S., De Carvalho F.A.T, Lurdermir T.B. (2006) *Hybrid Approach for interval-valued time Series Forecasting* Neurocomputing, Volume 71, Issues 16-18, October 2008, Pages 3344-3352
- [478] Makarenko A.V. (2011) *Phenomenological Model for Growth of Volumes Digital Data* Working Paper
- [479] Makridakis, S. (1989) Why combining works? *International Journal of Forecasting* 5, 601–603.
- [480] Makridakis, S. and Wheelwright, S.C. and Hyndman, R.J. (2008) *Forecasting methods and applications*, Wiley-India
- [481] Malkiel B.G. (1973) *A Random Walk Down Wall Street* New York: W. W. Norton & Co
- [482] Mandelbrot B. (1963) The Variation of Certain Speculative Prices *J. Business* XXXVI 392417
- [483] Mandelbrot B. Hudson R.L. (2006) *The Misbehavior of Markets: A Fractal View of Financial Turbulence* Basic Books; annotated edition edition
- [484] Manovich, L. (2011) *Trending: The Promises and the Challenges of Big Social Data*, Debates in the Digital Humanities, ed M.K.Gold. The University of Minnesota Press, Minneapolis, MN (15 July 2011) <http://www.manovich.net>
- [485] Mantegna R., Stanley H.E.(2000) *An Introduction to Econophysics* Cambridge University Press

- [486] Marchitelli M. 2009 *Metodi Statistici per la Valutazione Ex-Ante di Impatto della Regolamentazione* Ph.D. thesis, Dep. of Mathematics and Statistics, University Federico II Naples, in Italian.
- [487] Marron, J.S. (1987) A Comparison of Cross-Validation Techniques in Density Estimation *Annals of Statistics*, 15, 152-162
- [488] Marschinski, R. & Matassini, L. (2001) "Financial markets as a complex system: A short time scale perspective. *Research Notes* 01-4 Deutsche Bank Research.
- [489] Marvasti M. (2011) *Quantifying Information Loss Through Data Aggregation* Technical White Paper
- [490] Mason H. (2011) *What Data Tells Us* O'Reilly Strata Conference February 1-3 2011 Santa Clara US
- [491] Maté C. (2009) *The Bayesian approaches to combining forecasts. Some comments for inflation forecasting* Wienerwaldhof Workshop in Symbolic Data Analysis 2009
- [492] Maté C. (2011) "A multivariate analysis approach to forecasts combination. Application to Foreign Exchange (FX) markets", *Revista Colombiana de Estadística* vol. 34, no. 2, pp. 347-275, June 2011.
- [493] Maté C., Arroyo J. (2006), *Descriptive statistics for boxplot variables*, COMPSTAT 2006: 17th Conference of Int. Association for Statistical Computing and International Federation of Classification Societies 2006 Conference: Data Science and Classification. pp. 1549-1556. ISBN:3-7908-1708-2. Roma, Italy, 28 August- 1 September 2006
- [494] Maté C. Garcí Ascanio C. (2010) Electricity spot price forecasting using interval time series: A comparison between VAR and

- iMLP, 30th, International Symposium on Forecasting - ISF2010. San Diego, USA., 20-23 June 2010
- [495] Matei M. (2011) Non linear Volatility Modeling of Economic and Financial Time Series Using High Frequency Data. *Romanian Journal of Economic Forecasting* 2 2011
- [496] Mauboussin M.M., (2002). "Revisiting Market Efficiency: The Stock Market As A Complex Adaptive System," *Journal of Applied Corporate Finance* Morgan Stanley, vol. 14(4), pages 47-55.
- [497] McGroarty, F., Gwilym, O., Thomas, S., (2006). Microstructure effects, bid ask spreads and volatility in the spot foreign exchange market pre and post EMU, *Global Finance Journal* 17, 23.49
- [498] McKee, Gregory J. and Miljkovic, Dragan (2007) *Data Aggregation and Information Loss* 2007 Annual Meeting, July 29-August 1, 2007, Portland, Oregon TN 9843, American Agricultural Economics Association (New Name 2008: Agricultural and Applied Economics Association).
- [499] McKinsey (2011) *Big Data: the next frontier for innovation, competition and productivity* Report May 2011
- [500] Meijer E., Gilbert P.D. (2005) *Time Series Factor Analysis with an Application to Measuring Money* SOM Research Report, University of Groningen.
- [501] Meila, M. (2003) *Comparing Clusterings*. Working Paper COLT 2003.
- [502] Melnykov V. Maitra R. (2010) *Finite Mixture Models and Model Based Clustering* Statistics Surveys Vol.4, 80–116.

- [503] Melvin, M. and Yin, X., (2000). Public information arrival, exchange rate volatility, and quote frequency. *The Economic Journal*, 110(465), p.644661.
- [504] Milani F. (2008) *Analisi esplorativa dei dati: differenti metodi di rappresentazione grafica a confronto* Tesi di Laurea, Politecnico di Milano Facoltà di Ingegneria dei Sistemi Corso di Studi in Ingegneria Matematica
- [505] Miller G.A. (1956) *The magical number seven, plus or minus two: some limits on our capacity for processing information* Psychological Review 63 (2): 81–97.
- [506] Miller S. (2010) *Data, Data Everywhere* Information Management Blogs, March 8, 2010
- [507] Mineo A.M., Romito F. (2007) A method to clean up ultra high-frequency data, *Statistica Applicazioni*, 5, 167–186.
- [508] Mineo A. Romito F. (2008) *Different Methods to Clean Up Ultra High-Frequency Data* Atti della SIS Società Italiana di Statistica
- [509] Mingxing M., Jinwen W., Yonghong Z. (2005) *Application of Autoregressive Model to Monthly Runoff Probability Forecast* Dam Observation and Geotechnical Tests 2005-0
- [510] Mitsa T. (2010) Temporal Data Mining. *Clinics in Laboratory Medicine* Vol. 28. Chapman & Hall CRC.
- [511] Modugno M. (2011) *Nowcasting inflation using high frequency data* No.1324 European Central Bank Working Paper
- [512] Moore R. E. (1962) *Interval Arithmetic and Automatic Error Analysis in Digital Computing* Ph.D. Dissertation, Department of Mathematics, Stanford University, Stanford, California, Nov.

1962. Published as Applied Mathematics and Statistics Laboratories Technical Report No. 25.
- [513] Moore. R.E. (1966) *Interval Analysis* Prentice-Hall, Englewood Cliffs N. J., 1966.
- [514] Moore R. E.(1979): *Methods and Applications of Interval Analysis* Philadelphia SIAM
- [515] Mörchén, F. (2006) *Time Series Knowledge Mining* Phd Thesis, Philipps-University Marburg, Germany, Görich & Weiershäuser, Marburg, Germany, (2006), pp. 180 ISBN 3-89703-670-3
- [516] Mörchén F. (2006) *A better tool than allens relations for expressing temporal knowledge in interval data*. In T. Li, C. Perng, H. Wang, and C. Domeniconi, editors, Workshop on Temporal Data Mining at the Twelveth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2534, 2006
- [517] Mörchén F. (2007) *Unsupervised pattern mining from symbolic temporal data*. SIGKDD Explorations, 9:4155, 2007.
- [518] Mörchén F. (2011): *Temporal pattern mining in symbolic time point and time interval data* In Tutorial, SIAM International Conference on Data Mining.
- [519] Muchnik, L., Bunde, A. & Havlin, S. (2009) Long term memory in extreme returns of financial time series. *Physica A: Statistical Mechanics and its Applications* 388(19), 4145-4150.
- [520] Muller, U. A. Dacorogna, M., Olsen, R., Pictet, O. V. Schwarz, M. and Morgenegg, C. (1990) *Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis* Journal of Banking Finance, vol. 14, 6, 1189-1208.

- [521] Muller, U.A. (1996) *Volatility Computed by Time Series Operators at High Frequency* Working Paper presented at the Hong Kong International on Statistics in Finance 5-9 July 1999
- [522] Muñoz A., Maté C., Arroyo J., Sarabia A. (2007) "iMLP: Applying multi-layer perceptrons to interval-valued data", *Neural Processing Letters*. vol. 25, no. 2, pp. 157-169, April 2007.
- [523] Muñoz M.P., Corchero C. and Javier Heredia J.F. (2009) *Improving electricity market price scenarios by means of forecasting factor models* Dept. of Statistics and Operations Research Universitat Politècnica de Catalunya, Barcelona (Spain) DR 2009/6 1 June 2009
- [524] Murtagh F. (1985). A Survey of Algorithms for Contiguity-Constrained Clustering and Related Problems. *The Computer Journal*, 28(1), 82–88.
- [525] Mykland A. Zhang L. (2009) *The Econometrics of High Frequency Data* Working Paper February 22, 2009
- [526] Nadaraya E.A. (1964) On estimating regression, *Theory of probability and its applications* 9(1), 141–142.
- [527] Nagabhushan P. and Pradeep Kumar R. (2007) *Histogram PCA* International Symposium on Neural Networks - ISNN (2) 2007; Nanjing, China; pp1012–1021;
- [528] Nakamura T. Small M. (2007) *Tests of Random Walk Hypothesis for Financial Data* Physica A 377 (2007) 599-615
- [529] Nakano J., Fukui A. Shimizu (2011) Principal Component Analysis for Aggregated Symbolic Data. Workshop in Symbolic Data Analysis Namur, Belgium, June 2011

- [530] McNicholas P. (2007) *Model-Based Clustering: An Overview* Presentation
- [531] Moon, Y.I. and Rajagopalan, B. and Lall, U. (1995) Estimation of mutual information using kernel density estimators. *Physical Review E* 52,3, APS.
- [532] Nelsen R.B. (1999) *An Introduction to Copulas* Springer, New York.
- [533] Neumaier A. (1990) *Interval methods for systems of equations*, CUP, 1990
- [534] Newbold, P. and Granger, C. W. J. (1974) Experience with forecasting univariate time series and the combination of forecasts. *J. R. Statist. Soc. A* 137, 131–165.
- [535] Ng L. W. (2006) *Frequencies in Ultra-high-frequent Trading* Working Paper
- [536] Ng L. W. (2008) *Spectral Densities of Ultra-High Frequency Data* Centre for Computational Finance and Economic Agents Working Paper Series
- [537] Nguyen Quoc Viet Hung, Duong Tuan Anh (2007) Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Series. Proceedings of the 2007 IEEE International Symposium on Information Technology Convergence (ISITC 2007). IEEE Press.
- [538] Nickel K.(1969) *Triplex Algol and its Applications* in Topics in Interval Analysis E. Hansen ed. Oxford University Press Oxford 1969

- [539] Noirhomme–Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. Statistical Analysis and Data Mining (in press).
- [540] Orcutt, G.H. and Watts, H.W. and Edwards, J.B. (1968) Data aggregation and information loss *The American Economic Review* 58,4,773–787, 1968
- [541] O'Reilly *Strata Conference* <http://strataconf.com/strata2012>
- [542] Pagan, A.P. and A. Ullah (1999), *Nonparametric Econometrics* Cambridge University Press
- [543] Palumbo F. and Lauro C.N.(2003) A PCA for interval valued data based on midpoints and radii, in *New developments in Psychometrics*, Yanai H. et al. eds., Psychometric Society, Springer-Verlag, Tokyo.
- [544] Palumbo F. (2011) *Exploratory analysis for interval valued data* Keynote Lecture Cladag 2011 Pavia
- [545] Palumbo F. Gettler Summa M. (2000) Symbolic Interpretation of PCA on Interval Data, Atti della XL Riunione Scientifica della Società Italiana di Statistica Sessioni Plenarie e specializzate, pp. 103-114, Firenze 26-28 Aprile 2000.
- [546] Palumbo F. Marino M. (2002) Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression, *Statistica Applicata*, Vol. 14, 3, (2002) pp.277-291.
- [547] Palumbo F. Romano R. Esposito Vinzi V. (2008) "Fuzzy PLS Path Modeling: A New Tool for Handling Sensory Data" Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis and Knowledge Organization, 2008, XI, 689–696

- [548] Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business* 53, pp.61-65
- [549] Parzen E. (1962), On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* 33, 1065-1076.
- [550] Pasley A., Austin J. (2004) *Distribution forecasting of high frequency time series* Decis. Support Syst. 37, 4 (September 2004), 501-513
- [551] Pattarin, F. and Paterlini, S. and Minerva, T. (2004) Clustering financial time series: an application to mutual funds style analysis *Computational statistics & data analysis*, 47, 2, 353-372, 2004, Elsevier.
- [552] Patterson S. Rogow G. (2009) *What's Behind High-Frequency Trading*. The Wall Street Journal Money August 2 2009
- [553] Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 186: 343-326.
- [554] Pekalska, E. Duin R. P.W., & Paclik, P. (2006). Prototype Selection for Dissimilarity-Based Classifiers. *Pattern Recognition*, 39:2, pp. 189-208.
- [555] Peracchi F. (2001) *Econometrics* Wiley & Sons
- [556] Percival D.B. Walden A.T (2006) *Wavelet Methods for Time Series Analysis* Cambridge University Press

- [557] Pesaran M. H., and Pick A. (2010) Forecast Combination across Estimation Windows. *Journal of Business and Economic Statistics*, forthcoming
- [558] Pesaran, M.H., Timmermann, A. (2004) *Real Time Econometrics* IZA Discussion Papers 1108, Institute for the Study of Labor (IZA).
- [559] Pesaran, M. H., and Timmermann. A. (2007) "Selection of Estimation Window in the Presence of Breaks" *Journal of Econometrics* 137, 134-161.
- [560] Peters E.E. (1996) *Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility*, Second Edition, John Wiley and Sons
- [561] Piccinato B. Ballochi G. Dacorogna M.M. Gencçay R. (1999) Intraday Statistical Properties of Eurofutures. *Derivatives Quarterly*, Volume 6, Number 2, Winter 1999, A Publication of institutional Investor, Inc., 488 Madison Avenue, New York, N. Y. 10022, pp. 28-44
- [562] Piccolo D. (2000) *Statistica* Il Mulino
- [563] Pilar M. Muñoz M. Marquez D., Chulia H. (2008) *The Financial Crisis of 2008: Modelling the Transmission Mechanism Between the Markets* Working Paper presented at Compstat 2010 Paris
- [564] Potter C. (2006) *Visualization of Statistical Uncertainty* Working Paper
- [565] Potter K. and Kniss J. and Riesenfeld R. (2007) *Visual Summary Statistics*. Technical Report, Univeristy of Utah, no. UUCS-07-004.

- [566] Povinelli R.J. (2000) Identifying Temporal Patterns for Characterization and Prediction of Financial Time Series Events. Proceedings, pages 46–61, Springer
- [567] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [568] Rabinovich S.G. (1995) *Measurement errors and uncertainties: theory and practice* Springer
- [569] Racine J.S. (2008) *Nonparametric Econometrics: A Primer*. Foundation and Trends in Econometrics Vol. 3, No. 1 (2008) 1-88.
- [570] Refaeilzadeh P., Tang L. and Liu. H. 2009 *Cross Validation* In Encyclopedia of Database Systems, Editors: M. Tamer Azsu and Ling Liu. Springer, 2009.
- [571] Raftery A., Fraley C. (2007) *Model-based Methods of Classification: Using the mclust Software in Chemometrics* Journal of Statistical Software, American Statistical Association, vol. 18(i06).
- [572] Rajaraman, A. and Ullman D.J. (2010) *Mining of Massive Datasets* Lecture Notes for Stanford CS345A Web Mining: 328. <http://infolab.stanford.edu/ullman/mmds.html>.
- [573] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis* New York: Springer-Verlag.
- [574] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. New York: Springer-Verlag.

- [575] Rand W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* American Statistical Association 66 (336): 846850
- [576] Raykar V.C. (2007) *Computational tractability of machine learning algorithms for tall fat data* Working Paper.
- [577] Ras Z.W. Tsumoto S. Zighed D.A. (2005) *Complex Data Types* Proceedings of a Workshop held in Conjunction with 2005 IEEE International Conference on Data Mining, Houston, Texas, USA, November 27, 2005
- [578] Raudys S., 1991. On the effectiveness of Parzen window classifier. *Informatica* 2 (3), 434454.
- [579] R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0
- [580] Resti A. and Sironi A. (2007) *Risk Management and Shareholders Value in Banking*, John Wiley, New York.
- [581] Revol N. (2009) *Certified linear algebra Introduction to Interval Arithmetic* Lecture Notes Cours de recherche master informatique 19 October 2009
- [582] Riani M., (2004) *Extensions of the Forward Search to Time Series in Linear and Nonlinear Dynamics in Time Series* Estella Bee Dagum and Tommaso Proietti Editors, Vol.8 Issue 2 Article 2 Studies in Nonlinear Dynamics & Econometrics
- [583] Ricci V., (2006) *Principali tecniche di regressione con R* Versione 0.3 11 settembre 2006

- [584] Ritter G.X., Wilson J.N. (1996) *Handbook of Computer Vision Algorithms in Image Algebra* CRC Press, first edition
- [585] Rodríguez O. (2000) *Classification et Modèles Linéaires en Analyse des Données Symboliques*, Thèse de doctorat, Université Paris IX Dauphine.
- [586] Rodríguez O. (2004) *The Knowledge Mining Suite (KMS)*. Working Paper
- [587] Rodríguez O., Diday E., Winsberg S., (2000) *Generalizations of Principal Components Analysis* Working Paper
- [588] Rodríguez O., Diday E., Winsberg S., (2000) *Generalization of the Principal Components Analysis to Histogram Data* presented to the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, Lyon, France
- [589] Rodríguez O, Pacheco A. (2004). Applications of Histogram Principal Components Analysis, The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Vol. ECML/PKDD 2004 The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD): Pisa Italia.
- [590] Rodríguez O. Pacheco A. (2008) *Applications of Histogram Principal Components Analysis* Working Paper
- [591] Rodriguez P.M.M. Salish N. (2011) *Modeling and Forecasting Interval Time Series with Threshold Models: An Application to S&P500 Index Returns*. Working Paper Bank of Portugal.

- [592] Rogers, L.C.G. and Satchell, S.E., (1991) Estimating variance from high, low and closing prices. *The Annals of Applied Probability* 1, pp. 504-512.
- [593] Rokne J.G. (2001) *Interval arithmetic and interval analysis: an introduction* In Granular computing, Witold Pedrycz (Ed.). Physica-Verlag GmbH, Heidelberg, Germany, Germany 1-22
- [594] Romano E. Giordano G. Lauro C. (2006) *An Inter-Models Distance for Clustering Utility Functions* Statistica Applicata-Italian Journal of Applied Statistics, Vol.17, n.2.
- [595] Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* 27, 832837.
- [596] Ross S. M. (1996) Stochastic processes. Wiley, New York
- [597] Rosu I. (2009) A dynamic model of the limit order book, *Review of Financial Studies*, 22 (2009), pp. 4601–4641
- [598] Rousseeuw, P. J., van Driessen, K. (2006). Computing LTS Regression for Large Data Sets *Data Mining and Knowledge Discovery* 12, 29-45.
- [599] Ruppert D. (2010) *Statistics and Data Analysis for Financial Engineering* (Springer Texts in Statistics) Springer
- [600] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90, 12571270
- [601] Saita F. (2007), *Value at Risk and Bank Capital Management. Risk Adjusted Performances, Capital Management and Capital*

- Allocation Decision Making* Elsevier Academic Press, Advanced Finance Series, Burlington, MA;
- [602] Salish, N., and Rodrigues, P.M.M., (2010). *Non-Linearities in Interval Time Series* Working paper.
- [603] Salish N. Rodrigues P.M.M. (2010) *Interval time series models and forecast evaluation methods: An application to S&P500* Working Paper
- [604] Samworth, R. J., & Wand, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Annals of Statistics*, 38(3), 1767-1792. Retrieved from <http://arxiv.org/abs/1010.0591>
- [605] Samuelson P.A., (1965) *Proofs that Properly Anticipated Prices Fluctuate Randomly* Industrial Management Rev.6,41-45
- [606] Sánchez Úbeda E.F. (1999) *Modelos para el análisis de datos: contribuciones al aprendizaje a partir de ejemplos*. Thesis Doctoral
- [607] Saporta G. (1990) *Probabilités, Analyse des Données et Statistiques*. Edit. Technip Paris.
- [608] Schmidt, M., Lipson, H., (2008) "Coevolution of Fitness Predictors," IEEE Transactions on Evolutionary Computation, Vol.12, No.6, pp. 736-749.
- [609] Schmidt M., Lipson H. (2008), "Data-mining Dynamical Systems: Automated Symbolic System Identification for Exploratory Analysis", Proceedings of the 9th Biennial ASME Conference on Engineering Systems Design and Analysis (ESDA08), Haifa, Israel, July 7-9, 2008.

- [610] Schmidt M., Lipson H. (2007), "Comparison of Tree and Graph Encodings as Function of Problem Complexity", Genetic and Evolutionary Computation Conference (GECCO'07), pp. 1674-1679.
- [611] Schmidt M., Lipson H. (2009), "Symbolic Regression of Implicit Equations," Genetic Programming Theory and Practice, Vol. 7, Chapter 5, pp. 73-85.
- [612] Schmidt M., Lipson H. (2009) "Incorporating Expert Knowledge in Evolutionary Search: A Study of Seeding Methods," Genetic and Evolutionary Computation Conference (GECCO'09).
- [613] Schmidt M., Lipson H. (2009) "Solving Iterated Functions Using Genetic Programming," Genetic and Evolutionary Computation Conference, Late Breaking Paper (GECCO'09).
- [614] Schmidt M., Lipson H. (2007), "Learning Noise", Genetic and Evolutionary Computation Conference (GECCO'07), pp. 1680-1685.
- [615] Schweizer B. (1984). *Distributions are the Numbers of the Future*. Proceedings The Mathematics of Fuzzy Systems Meeting University of Naples, 137-149.
- [616] Science (2011) *Dealing with Data Science* Special Issue February 11 2011
- [617] Scott D. W. (1979). *On optimal and data-based histograms*. Biometrika, 66, 605-610.
- [618] Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

- [619] Sewell M. (2008) *Characterization of Financial Time Series* Working Paper
- [620] Sewell M. (2008) *Chaos in Financial Markets* Working Paper
Department of Computer Science University College London
- [621] Sewell M. (2011) *Characterization of Financial Time Series*
UCL Research Note RN/11/01 January 2011
- [622] Sheather, S.J. (1992). The performance of six popular bandwidth selection methods on some real data sets (with discussion). *Computational Statistics* 7: 225-250, 271-281.
- [623] Sheather, S, and J S Marron. 1990. Kernel quantile estimators. *Journal of the American Statistical Association* 85, no. 410: 410-416. <http://www.jstor.org/stable/2289777>.
- [624] Sheather, S. J. and Jones, M. C. (1991) *A reliable data-based bandwidth selection method for kernel density estimation* JRSS-B 53, 683-690.
- [625] Shen H., and Huang Z.J. (2008) Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics* 2, no. 2: 601-623.
- [626] Shiller R.J. (2005) *Irrational Exuberance* 2nd edition, Princeton University Press 2005 datasets:
<http://www.econ.yale.edu/shiller/data.htm>
- [627] Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51, 492-508.
- [628] Shimodaira, H. (2004) Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, 32, 2616-2641.

- [629] Shumway R. H. Stoffer D.S. 2011 *Time Series Analysis and Its Applications* Springer Texts in Statistics 3rd ed.
- [630] Signoriello S. (2008) *Contributions to Symbolic Data Analysis: A Model Data Approach* Ph.D Thesis in Statistics, University Federico II of Naples
- [631] Silverman B.W., (1978). Choosing the window width when estimating a density. *Biometrika* 65, 111.
- [632] Silverman, B.W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society, Series B* 43:97-99
- [633] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis* Chapman and Hall, London.
- [634] Simonoff, J. S. (1996) *Smoothing Methods in Statistics* New York: Springer
- [635] Smith, J, and Wallis K.F. (2009). A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics* 71, no. 3: 331-355.
- [636] Sornette D. 2004 *Why Stock Markets Crash: Critical Events in Complex Financial Systems* Princeton University Press
- [637] Springer, M.D. (1979) *The Algebra of Random Variables* Wiley.
- [638] Spurgin, R B and Schneeweis, T (1999). *Efficient Estimation of Intra-day Volatility: A Method-of-Moments Approach Incorporating Trading Range*, in Lequeux, P (ed.), *Financial Markets Tick by Tick*, London: John Wiley and Sons

- [639] Sreedharan J.N. (2004) *A Vector Error Correction Model (VECM) of Stockmarket Returns*. Working Paper
- [640] Starica C. and Granger C. (2005) Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87(3):503-522, 09 2005.
- [641] Stéphán V., Hébrail G. Lechevallier Y. *Generation of Symbolic Objects from Relational Databases*
- [642] Sterne, J. A. C., Davey-Smith, G., and Cox, D. R. (2001), Sifting the Evidence: Whats Wrong with Significance Tests? *British Medical Journal* 322, 226–231
- [643] Strehl A. and Ghosh J. (2002) Cluster ensembles a knowledge reuse framework for combining multiple partitions, *Journal on Machine Learning Research (JMLR)* 2002
- [644] Stock J.H Watson M.W. (2004) *Forecasting with Many Predictors* Handbook of Economic Forecasting
- [645] Sturges H. A. (1926) The choice of a class interval. *Journal Of The American Statistical Association* 21: 65-66
- [646] Sun, W., Rachev, S. T., and Fabozzi, F. (2007). *Long-range dependence, fractal processes, and intra-daily data*. In F.Schlottmann, D. Seese,& C.Weinhardt (Eds.), Handbook of IT and Finance. Heidelberg: Springer Verlag.
- [647] Sunaga T. (1958), *Theory of interval algebra and its application to numerical analysis* In: Research Association of Applied Geometry (RAAG) Memoirs, Ggujutsu Bunken Fukuy-kai. Tokyo, Japan, 1958, Vol. 2, pp. 29-46 (547-564); reprinted in Japan Journal on Industrial and Applied Mathematics, 2009, Vol. 26, No. 2-3, pp. 126-143.

- [648] Sun Y. Genton M. G. (2009) *Functional Boxplots for Complex Data Visualizations* Working Paper
- [649] Suzuki, R. and Shimodaira, H. (2004) *An application of multi-scale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?* The Fifteenth International Conference on Genome Informatics 2004, P034.
- [650] Tam, K.Y., Kiang M.Y, and Chi R.T. (1991) "Inducing Stock Screening Rules for Portfolio Construction," *The Journal of the Operational Research Society*, Vol. 42, No. 9:747-757, 1991.
- [651] Tarantelli E. (1988) *L'utopia dei deboli è la paura dei forti. Saggi, relazioni e altri scritti accademici* Istituto di politica economica della Facoltà di economia e commercio dell'Università di Roma 1a edizione 1988
- [652] Taschini S. *Interval Arithmetic: Python Implementation and Applications* in Proceedings of the 7th Python in Science conference (SciPy 2008), G Varoquaux, T Vaught, J Millman (Eds.), pp. 16-2;
- [653] Tay, A. S., & Wallis, K. F. (2000). Density forecasting: a survey. *Journal of Forecasting*, 19(4), 235-254.
- [654] Teles P.and Brito M.P. (2005) *Modelling interval time series data* In 3rd IASC world conference on Computational Statistics Data Analysis, Limassol, Cyprus, 2005
- [655] The Economist (2010) *A special report on managing information: Data, data everywhere* February 25 2010
- [656] The Economist (2011) *Building with big data* Printed Edition May 26th 2011

- [657] Timmermann A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, 19(4), p.231-234
- [658] Timmermann A. (2006) Forecast Combinations, *Handbook of Economic Forecasting*, pp.135-196.
- [659] Tiozzo L. (2011) *Market Microstructure and High frequency data: Is Market efficiency still a reasonable hypothesis? A survey* Working Paper
- [660] Titterton, D.M., Smith A.F.M. and Makov U.E. (1985) *Statistical Analysis of Finite Mixture Distributions* Wiley, New York. x+243 pp.
- [661] Tjung, L.C., Kwon, O., Tseng K. C., Bradley-Geist J. (2010) Forecasting Financial Stocks using Data Mining. *Global Economy and Finance Journal* Volume 3. Number 2. September 2010. Pp. 13 - 26
- [662] Toit, S. H. and Browne, M. W. (2007). Structural equation modeling of multivariate time series. *Multivariate Behavioral Research*, 42(1):67–101.
- [663] Tong, H. & Lim, K. S. (1980) Threshold Autoregression, Limit Cycles and Cyclical Data (with discussion), *Journal of the Royal Statistical Society, Series B*, 42, 245–292.
- [664] Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis* Lecture Notes in Statistics, Springer-Verlag.
- [665] Tong, H. (1990) *Non-Linear Time Series: A Dynamical System Approach* Oxford University Press.

- [666] Toulemonde G., Guillou A., Naveau P., Vrac M., Chevallier F.. (2009) "Autoregressive models for maxima and their applications to CH₄ and N₂O", *Environmetrics*, in press, 2009 (preprint).
- [667] Tsay R.S. (2005) *Analysis of Financial Time Series* Wiley & Sons
- [668] Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press
- [669] Turlach, B. A. (1993) Bandwidth Selection in Kernel Density Estimation: A Review. CORE and Institut de Statistique 19, no. 4: 1-33.
- [670] Tukey J.W. (1977) *Exploratory Data Analysis* Addison-Wesley.
- [671] Tzitzikas Y. (2004) *An Algebraic Method for Compressing Very Large Symbolic Data Tables* Workshop on Symbolic and Spatial Data Analysis (SSDA 2004) of ECML/PKDD 2004, Pisa, Italy, September 2004
- [672] Van Der Laan M., Hsu J.P. and Peace K.E. (2010) *Next Generation of Statisticians Must Build Tools for Massive Data Sets*, Amstat News 1 september 2010
- [673] Valova I. Noirhomme Fraiture M. (2008) *Processing of Large Data Sets: Evolution, Opportunities and Challenge* Proceedings of PCaPAC08, Ljubljana, Slovenia.
- [674] Vance A. (2010) *Start-Up Goes After Big Data With Hadoop Helper*. New York Times Blog. April 22, 2010. <http://bits.blogs.nytimes.com/2010/04/22/start-up-goes-after-big-data-with-hadoop-helper/?dbk>

- [675] Venables, W. N., Ripley, B. D. (2002) *Modern Applied Statistics with S* Fourth Edition. Springer, New York.
- [676] Verde R., Irpino A. (2006). *A new Wasserstein based distance for hierarchical clustering of histogram symbolic data* In: Batagelj V., Bock H.-H., Ferligoj A., Ziberna A. (Eds). *Data Science and Classification*. pp. 185-192. ISBN: 3-540-34415-2. Berlin: Springer (Germany).
- [677] Verde R., Irpino A. (2008). *Ordinary Least Squares for Histogram Data based on Wasserstein Distance* COMPSTAT 2010 Paris –August 22-27
- [678] Verde R., Irpino A. (2011). *Basic statistics for probabilistic symbolic variables: a novel metric-based approach* Working Paper.
- [679] Vichi M., Rocci R., Kiers H.A.L. (2007) Simultaneous Component and Clustering models for three-way data: Within and Between Approaches. *Journal of Classification*, 24, 1, 71-98.
- [680] Viertl R. (2007): *Univariate Statistical Analysis with Fuzzy Data* Working Paper Vienna University of Technology
- [681] Vinzi, V. E., Lauro, C., and Amato, S. (2004). PLS typological regression: Algorithmic, classification and validation issues. In M. Vichi, P. Monari, S. Mignani, & A. Montanari(Eds.), *New developments in classification and data analysis* (pp. 133-140). Berlin et al.: Springer.
- [682] Vishnyakov B. V. Kibzun A. I. (2007) Application of the bootstrap method for estimation of the quantile function. *Automation and Remote Control* Volume 68, Number 11, 1931-1944
- [683] Very Large Data Base Endowment Inc (2010) website, version the 4 october 2010

- [684] Vogels W. (2011) *Data Without Limits* O'Reilly Strata Conference February 1-3 2011 Santa Clara US
- [685] Wagner, S., Wagner, D. (2006): *Comparing Clusterings An Overview*. Technical Report 2006-4, Faculty of Informatics, Universitat Karlsruhe (TH) (2006)
- [686] Wand, M. P. (1995). *Data-based choice of histogram binwidth*. The American Statistician, 51, 5964.
- [687] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.
- [688] Wang H., Guan R. and Wu J. (2011) *Linear Regression of Interval-valued Data based on Complete Information in Hypercubes* Workshop in Symbolic Data Analysis Namur, Belgium, June 2011
- [689] Wansbeek T. Meijer E. (2005) *Factor analysis (and beyond)* Faculty of Economics University of Groningen 1 March 2005 CES
- [690] Wansbeek T. and Meijer. E. (2000) *Measurement Error and Latent Variables in Econometrics*. North-Holland, Amsterdam.
- [691] Warmus M. (1956) Calculus of approximations, *Bull. Acad. Polon. Sci.*, Cl. III, IV (1956) 253-259.
- [692] Warmus M. (1961) Approximations and inequalities in the calculus of approximations: Classification of approximate numbers. *Bull. Acad. Polon. Sci., Ser. Math, Astr. et Phys.* IX (1961) 241-245.
- [693] Warner R.M. (1998) *Spectral analysis of time-series data* Guilford Press.

- [694] Watson G.S. (1964) Smooth regression analysis, *Sankhya: The Indian Journal of Statistics (Series A)*, 26(4), 359-372.
- [695] Weckman, G.R. et al., 2008. An integrated stock market forecasting model using neural networks. *International Journal of Business Forecasting and Marketing Intelligence*, 1(1), p.30
- [696] Weiss S.M. Indurkha N. (1997) *Predictive Data Mining: A Practical Guide* (The Morgan Kaufmann Series in Data Management Systems) Morgan Kaufmann, 1 edition
- [697] West K. (2006) *Forecast Evaluation* Handbook of Economic Forecasting, Volume 1 Edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann 2006 Elsevier B.V.
- [698] White T. (2009) *Hadoop: The Definitive Guide* 1st Edition. O'Reilly Media. Pg 3.
- [699] Wickham H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York.
- [700] Wildi M. (2007) *NN3-Forecasting Competition: an Adaptive Robustified Multi-Step-Ahead Out-Of-Sample Forecasting Combination* NN3 Forecasting Competition Working Paper
- [701] Williamson R.C. (1989) *Probabilistic Arithmetic* Ph.D Thesis Department of Electrical Engineering University of Queensland August 1989
- [702] Williamson, R.C., and T. Downs (1990) Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4:89-158.

- [703] Winkler R.L. and Makridakis S. (1983) The Combination of Forecasts *Journal of the Royal Statistical Society. Series A (General)* , Vol. 146, No. 2 (1983), pp. 150–157
- [704] Witten, I.H. and Frank, E., (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* Morgan Kaufmann
- [705] Wolfe, J.H. (1963) *Object Cluster of Social Areas* Master's Thesis University of California, Berkeley.
- [706] Xenomorph (2007) *High Frequency Data Analysis: Considered decision making with TimeScape* Xenomorph Working Paper
- [707] Xu S., Chen X. and Han, A. (2008). Interval forecasting of crude oil price. *Interval and Probabilistic Uncertainty*, 46, 353–363.
- [708] Xu, R. and Wunsch, D. and others (2005) Survey of clustering algorithms, *Neural Networks, IEEE Transactions on*,16,3,645–678,2005,IEEE
- [709] Yakowitz, S. (1987) Nearest-Neighbour Methods for Time Series Analysis. *Journal of Time Series Analysis* 8(2), 235-247.
- [710] Yan, B., Zivot, E. (2003): *Analysis of High-Frequency Financial Data with S-Plus* Working Paper, University of Washington.
- [711] Yang B. Zivot E. (2003) *Analysis of High-Frequency Financial Data with S-Plus* Working Paper
- [712] Young R.C.(1931): The algebra of many-valued quantities *Math. Ann*, 104, 260-290.
- [713] Zeileis A., Kleiber C., Kraemer W. and Hornik K. (2003). Testing and Dating of Structural Changes in Practice. *Computational Statistics & Data Analysis*, 44, 109-123.

- [714] Zeileis A. (2005) A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, 24(4):445–466
- [715] Zeileis A. (2006) Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis*, 50:2987–3008, 2006
- [716] Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and Neural Network model. *Neurocomputing* 50, 159-175.
- [717] Zhang Z. (2007) *Functional data analysis for densities* Ph.D. Thesis, University of California, Davis, 2007.
- [718] Zhang Q., Keasey K., and Cai C. X. (2009) Forecasting Using High-Frequency Data: A Comparison of Asymmetric Financial Duration Models. *Journal of Forecasting* Vol. 28, No. 5, pp. 371-386, August 2009. Available at SSRN: <http://ssrn.com/abstract=1687000>
- [719] Zhang X. King L.M. Hyndman R.J. (2004) *Bandwidth Selection for Multivariate Kernel Density Estimation Using MCMC* Working Paper
- [720] Zhang Z. Muller H.G. (2010) *Functional Density Synchronization* Working Paper
- [721] Zhou, B. (1996) *High-frequency data and volatility in foreign-exchange rates*, *Journal of Business and Economic Statistics* 14, 455-462.
- [722] Zivot E. (2005) *Analysis of High Frequency Financial Data: Models, Methods and Software. Part I: Descriptive Analysis of High Frequency Financial Data with S-PLUS* Working Paper

- [723] Zivot E. Yang J. (2006) *Modeling financial time series with S-PLUS* Vol.13 Birkhauser
- [724] Zou, H. and Y. Yang, 2004. Combining time series models for forecasting. *International Journal of Forecasting* 20, 69–84.
- [725] Zuccolotto P. (2011): Symbolic missing data imputation in principal component analysis. *Statistical Analysis and Data Mining* 4(2): 171-183 (2011)
- [726] Zumbach G.O. Muller U.A. (2001) *Operators on inhomogeneous time series* International Journal of Theoretical and Applied Finance, 4(1),147-178

Websites and Electronical Sources

- [727] A.A.V.V. *AIDA – Abstract Interfaces for Data Analysis*
<http://aida.freehep.org/> Web. 30 July 2011
- [728] A.A.V.V. *Comparing Partitions*
<http://darwin.phyloviz.net/ComparingPartitions/> Web. 30 July 2011
- [729] A.A.V.V. *Reproducible Research*
<http://reproducibleresearch.net/index.php/Mainpage> Web. 1 August 2011
- [730] A.A.V.V. *R Graph Gallery* <http://addictedtor.free.fr/graphiques/>
Web 9 August 2011
- [731] A.A.V.V. *R Graphical Manual*
<http://rgm2.lab.nig.ac.jp/RGM2/images.php?show=allpageID=1540>
Web 9 August 2011
- [732] A.A.V.V. *RAMAS Risk Calc version 4.0*
<http://www.ramas.com/riskcalc.htm> used for Web 22 August 2011

Bibliography

- [733] A.A.V.V. *RAMAS Constructor*
<http://www.ramas.com/constructor.htm> Web 22 August 2011
- [734] A.A.V.V. *Wolfram* <http://www.wolfram.com/> Web 19 september 2010
- [735] AIDA Abstract Interfaces for Data Analysis
<http://aida.freehep.org/>
- [736] Estimating the Hurst Exponent
http://www.bearcave.com/misl/misl_tech/wavelets/hurst/index.html Web 15 August 2011
- [737] Coheris Performance Driven Software (SPAD)
<http://www.coheris.fr/en/page/produits/SPAD-data-mining.html>
 Web 12 September 2011
- [738] CrossValidated <http://stats.stackexchange.com/> Web 15 August 2011
- [739] Fair R. (2011) Fairmodel <http://fairmodel.econ.yale.edu/> Web 31 October 2011
- [740] Intelligent Trading Blog Practical Implementation of Neural Network based time series (stock) prediction
<http://intelligenttradingtech.blogspot.com/2010/01/systems.html>
 Web 20 October 2011
- [741] A.A.V.V. *Scholarpedia* http://www.scholarpedia.org/article/Main_page Web 4 August 2011
- [742] A.A.V.V. *SYROKKO, éditeur innovant de logiciels de datamining*
<http://syrokko.com/index.php?page=home> Web. 5 Aug. 2011
- [743] A.A.V.V. *ASSO Project Analysis System of Symbolic Official data*
<http://www.info.fundp.ac.be/asso/index.html> Web. 5 Aug. 2011

- [744] A.A.V.V. *Wikinvest* <https://www.wikinvest.com/?type=classic> Web 14 August 2011
- [745] A.A.V.V. *Wikipedia* <http://www.wikipedia.org/> Web. 30 July 2011
- [746] Browne T. *High frequency data series cleaning in R* <http://stats.stackexchange.com/questions/11531/high-frequency-data-series-cleaning-in-r> Crossvalidated Web. 7 Aug. 2011
- [747] Cornell Creative Machines Lab *Eureqa* <http://creativemachines.cornell.edu/eureqa> Web 24 Aug. 2011
- [748] Dowe D. *Mixture Modelling page* <http://www.csse.monash.edu.au/dld/mixture.modelling.page.html> Web 31 July 2011
- [749] Drago C. *Carlo Drago's Bookmarks* <http://www.delicious.com/c.drago> Web. 6 Aug. 2011
- [750] Economic Research Federal Reserve Bank of St. Louis <http://research.stlouisfed.org/> Web 15 Aug. 2011
- [751] Glynn E.A. (2005) *Correlation "Distances" and Hierarchical Clustering* Stowers Institute for Medical Research 29 Dec. 2005
- [752] Granville V. <http://www.analyticbridge.com/>
- [753] Hall M., Frank E., Geoffrey Holmes G., Pfahringer B., Reutemann P., Witten I.H. (2009); *The WEKA Data Mining Software: An Update* SIGKDD Explorations, Volume 11, Issue 1.
- [754] Hyndman R.J. <http://robjhyndman.com/>
- [755] Kreinovich V. *Interval Computations* <http://www.cs.utep.edu/interval-comp/> Web 25 October 2011

- [756] Keogh E. (2011) *SAX Symbolic Aggregate approximation* Web 17 Aug. 2011
- [757] *MIX Software for Mixture Distributions* <http://www.math.mcmaster.ca/peter/mix/mix.html> Web. 31 July 2011
- [758] *PAST Paleontological Statistics software* <http://folk.uio.no/ohammer/past/> Web 30 August 2011
- [759] Rodríguez O. Home Page. <http://www.oldemarrodriguez.com/> Web 28 October 2011
- [760] *Stata 12* StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP. Web August 30 August 2011
- [761] Martin Sewell Finance <http://finance.martinsewell.com/> Web 23 October 2011
- [762] Shimazaki H. *Kernel Bandwidth Optimization* <http://toyoizumilab.brain.riken.jp/hideaki/res/kernel.html>
- [763] Suzuki R. and Shimodaira H. *An R package for hierarchical clustering with p-values* <http://www.is.titech.ac.jp/shimo/prog/pvclust/usage> Web. 31 July 2011
- [764] Ryan J.A *Quantmod: Quantitative Financial Modelling Trading Framework for R* <http://www.quantmod.com/> Web 13 August 2011
- [765] Oracle *Virtualbox* <http://www.virtualbox.org/> Web 24 August 2011
- [766] Vistocco D. *Two histograms in the same graph* <https://stat.ethz.ch/pipermail/r-help/2008-January/152614.html>
- [767] Wallis, K.F., 1989. Macroeconomic Forecasting - A Survey *The Economic Journal*, 99(394), p.28-61.

- [768] Wang Q. Megalooikonomou V. (2008), A dimensionality reduction technique for efficient time series similarity analysis *Inf. Syst.* 33, 1 (Mar.2008), 115- 132.
- [769] Warrenliao T. 2005. Clustering of time series dataa survey. *Pattern Recognition* 38, no. 11: 1857-1874.
- [770] Wickham H. (2011) *ggplot2* <http://had.co.nz/ggplot2/> Web 9 Aug. 2011
- [771] Wicklin R. (2011) *The Area under a Density Estimate Curve: Nonparametric Estimates* The Do loop SAS <http://blogs.sas.com/iml/index.php?/archives/157-The-Area-under-a-Density-Estimate-Curve-Nonparametric-Estimates.html> Friday, July 8. 2011
- [772] Wilson D.R. Martinez T.R. (1997). Instance Pruning Techniques. *ICML 97* Morgan Kaufmann, pp. 403-411.
- [773] Yahoo Finance <http://finance.yahoo.com/> Web 15 Aug. 2011

Python Packages (in alphabetical order)

- [774] interval Interval Arithmetic in Python <http://pyinterval.googlecode.com/svn/trunk/html/index.html>
- [775] IntPy Interval Arithmetic package in Python <http://pypi.python.org/pypi/IntPy>
- [776] PaCAL - ProbAbilistic CALculator <http://pacal.sourceforge.net/index.html>
- [777] paida - PAIDA is pure Python scientific analysis package and supports AIDA (Abstract Interfaces for Data Analysis) <http://paida.sourceforge.net/>

[778] Python interval <http://code.google.com/p/python-interval/>

R Packages (in alphabetical order)

[779] Peter Kampstra (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, Code Snippets 28(1). 1-9. URL <http://www.jstatsoft.org/v28/c01/>.

[780] Aron Charles Eklund (2010). beeswarm: The bee swarm plot, an alternative to stripchart.. R package version 0.0.7. <http://CRAN.R-project.org/package=beeswarm>

[781] John Fox and Sanford Weisberg (2011). *An R Companion to Applied Regression*, Second Edition. Thousand Oaks CA: Sage. URL: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

[782] Venables, W. N. Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[783] Marek Walesiak Andrzej Dudek andrzej.dudek@ue.wroc.pl (2011). clusterSim: Searching for optimal clustering procedure for a data set. R package version 0.39-2. <http://CRAN.R-project.org/package=clusterSim>

[784] Clustering procedures of symbolic objects that are described by discrete distributions by Natasa Kejzar and Vladimir Batagelj 2011. <https://r-forge.r-project.org/projects/clamix/>

[785] Lukasz Nieweglowski. (2009). clv: Cluster Validation Techniques. R package version 0.3-2. <http://CRAN.R-project.org/package=clv>

[786] Toni Giorgino (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7), 1-24. URL <http://www.jstatsoft.org/v31/i07/>

- [787] Simon Urbanek and Yossi Rubner (2011). `emdist`: Earth Mover's Distance. R package version 0.2-1. <http://CRAN.R-project.org/package=emdist>
- [788] Diethelm Wuertz, many others and see the SOURCE file (2009). `fExtremes`: Rmetrics - Extreme Financial Market Data. R package version 2100.77. <http://CRAN.R-project.org/package=fExtremes>
- [789] Diethelm Wuertz, Yohan Chalabi with contribution from Michal Miklovic, Chris Boudt, Pierre Chausse and others (2009). `fGarch`: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R package version 2110.80. <http://CRAN.R-project.org/package=fGarch>
- [790] Rob J Hyndman (2011). `forecast`: Forecasting functions for time series. R package version 2.16. <http://CRAN.R-project.org/package=forecast>
- [791] Hennig C. (2010). `fpc`: Flexible procedures for clustering. R package version 2.0–3. <http://CRAN.R-project.org/package=fpc>
- [792] Mehmet Hakan Satman (2011). `galts`: Genetic algorithms and C-steps based LTS (Least Trimmed Squares) estimation. R package version 1.1. <http://CRAN.R-project.org/package=galts>
- [793] Wickham H. (2009) `ggplot2`: elegant graphics for data analysis. Springer New York.
- [794] Nathan Stephens and Vicky Yang. (2009). `hacks`: Convenient R Functions. R package version 0.1-9. <http://CRAN.R-project.org/package=hacks>
- [795] Rob J Hyndman with contributions from Jochen Einbeck and Matt Wand (2010). `hdrcde`: Highest density regions and condi-

- tional density estimation. R package version 2.15. <http://CRAN.R-project.org/package=hdrdce>
- [796] Frank E Harrell Jr and with contributions from many other users. (2010). Hmisc: Harrell Miscellaneous. R package version 3.8-3. <http://CRAN.R-project.org/package=Hmisc>
- [797] iRegression (2011) by Eufrasio de A. Lima Neto and Claudio A. Vasconcelos
- [798] Ricardo Jorge de Almeida Queiroz Filho and Roberta Andrade de Araujo Fagundes (2011). ISDA.R: interval symbolic data analysis for R. R package version 1.0. <http://CRAN.R-project.org/package=ISDA.R>
- [799] S original by Matt Wand. R port by Brian Ripley. (2010). KernSmooth: Functions for kernel smoothing for Wand Jones (1995). R package version 2.23-4. <http://CRAN.R-project.org/package=KernSmooth>
- [800] MAINT.DATA: Modeling and Analysing Interval Data in R by Pedro Duarte Silva and Paula Brito
- [801] Chris Fraley and Adrian E. Raftery (2006) MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical Report No. 504, Department of Statistics, University of Washington(revised 2009)
- [802] Peter Macdonald and with contributions from Juan Du (2010). mixdist: Finite Mixture Distribution Models. R package version 0.5-3. <http://CRAN.R-project.org/package=mixdist>
- [803] Peter Carl and Brian G. Peterson (2010). PerformanceAnalytics: Econometric tools for performance and risk analysis.. R package version 1.0.3.2.

- [804] Matt Shotwell (2010). `profdpm`: Profile Dirichlet Process Mixtures. R package version 2.0. <http://CRAN.R-project.org/package=profdpm>
- [805] Hidetoshi Shimodaira (2009). `pvclust`: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 1.2-1. <http://www.is.titech.ac.jp/~shimo/prog/pvclust/>
- [806] Jeffrey A. Ryan (2010). `quantmod`: Quantitative Financial Modelling Framework. R package version 0.3-15. <http://CRAN.R-project.org/package=quantmod>
- [807] Juggins, S., (2009). `rioja`: Analysis of Quaternary Science Data, R package version 0.5-6.
- [808] Anthony Atkinson, Andrea Cerioli and Marco Riani. (2005). `Rfwdmv`: Forward Search for Multivariate Data. R package version 0.72-2. <http://www.riani.it>
- [809] Martin Maechler and many others. (2010). `sfsmisc`: Utilities from Seminar fuer Statistik ETH Zurich. R package version 1.0-14. <http://CRAN.R-project.org/package=sfsmisc>
- [810] Achim Zeileis, Friedrich Leisch, Kurt Hornik and Christian Kleiber (2002). `strucchange`: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7(2), 1-38. URL <http://www.jstatsoft.org/v07/i02/>
- [811] Analysis of symbolic data by Andrzej Dudek 2010 <http://keii.ae.jgora.pl/symbolicDA/index.html>
- [812] Agustin Mayo Iscar, Luis Angel Garcia Escudero and Heinrich Fritz (2010). `tlust`: Robust Trimmed Clustering. R package version 1.0-8. <http://CRAN.R-project.org/package=tlust>

- [813] Gilbert, Paul D. and Meijer, Erik Time Series Factor Analysis with an Application to Measuring Money, Research Report 05F10, University of Groningen, SOM Research School. Available from <http://som.eldoc.ub.rug.nl/reports/themeF/2005/05F10/>
- [814] Daniel Adler (2005). vioplot: Violin plot. R package version 0.2. <http://wsopuppenkiste.wiso.uni-goettingen.de/~dadler>
- [815] Solomon Messing (2010). wvioplot: Weighted violin plot. R package version 0.1. <http://CRAN.R-project.org/package=wvioplot>
- [816] David B. Dahl (2009). xtable: Export tables to LaTeX or HTML. R package version 1.5-6. <http://CRAN.R-project.org/package=xtable>